"Anyone concerned with language, semantics, or categorization will want to have this encyclopedic collection."

Professor Eleanor Rosch, Dept of Psychology, University of California, Berkeley, USA
PREFACE

The content of most of the chapters in this book was presented as part of the Summer Institute on categorization that took place at the Université du Québec à Montréal (UQAM) for 10 days in June and July 2003. The objective of this Institute was to address the problem of categorization through the lens of all the disciplines that are at the heart of the Cognitive Sciences: Cognitive Anthropology, Linguistics, Philosophy, Neuroscience, Psychology and Cognitive Computer Science. This book is a natural, concrete outcome of the Institute.

In planning this book, we wanted all aspects of categorization to be represented. We therefore filled the holes in the original program by soliciting contributions from researchers who had not been involved directly in the Summer Institute on categorization. In its present form, the book contains some 50 chapters. To our knowledge, it is the first time in history that the problem of categorization has been considered from so many angles within a single book.

We would like to thank the members of the scientific committee of the Summer Institute for their contributions, which have had a major impact on the contents of this book (Henrietta Cedergren for Linguistics, Pierre Poirier for Philosophy, Henri Cohen for Neuroscience, Stevan Harnad for Psychology, Bernard Lefebvre for Cognitive Computer Science, and Claire Lefebvre as the Director of the Institute). We would like to thank Robert Proulx, Dean of the Faculty of Human Sciences, for his support throughout. The financial contribution from the Faculty enabled us to prepare this exceptionally large manuscript. Last but not least, we are grateful to Zofia Laubitz, Marlene Busko, and Sanja Obradović for making an important contribution by copy-editing and finalizing the manuscript.

Henri Cohen
Claire Lefebvre
This page intentionally left blank
TABLE OF CONTENTS

See Color Plate Section at the back of this book

Preface v

Bridging the Category Divide
HENRI COHEN and CLAIRE LEFEBVRE

1. Introduction 2
2. Organization of the book 2
3. Major common themes 7
   3.1. The notions of category and categorization 7
   3.2. The nature of categories: Discrete, vague, or other 9
   3.3. Are there modality effects on categories? 10
   3.4. Are there universal categories? Are there innate categories? 11
4. Bridging the category divide 13
References 15

PART 1 – CATEGORIZATION IN COGNITIVE SCIENCE

Chapter 1
To Cognize is to Categorize: Cognition is Categorization
STEVAN HARNAD
Abstract 20
1. Sensorimotor systems 21
2. Invariant sensorimotor features (“affordances”) 21
3. Categorization 21
4. Learning 22
5. Innate categories 23
6. Learned categories 23
7. Supervised learning 24
8. Instrumental (operant) learning 24
9. Color categories 25
10. Categorical perception 25
11. Learning algorithms 26
12. Unsupervised learning 27
13. Supervised learning 27
14. Vanishing intersections? 28
15. Direct sensorimotor invariants 29
Chapter 2
A Modular Approach to Grammatical Categories Evidence from Language Diversity and Contact
PIETER MUYSKEN

Abstract 46

1. Introduction 47
2. Modularity and mismatch 48
3. Grammaticalization: The case of pe in Sranan (Suriname) 52
4. Mismatches in complexity of representations: The case of ku in Cuzco Quechua (Peru) 54
5. Lexical nondistinctness 55
   5.1. The case of timap in Palikur 55
   5.2. ‘For’ prepositions to become complementizers 56
   5.3. Adjectives versus adverbs 57
6. Lexical overspecification: Dutch gender and definiteness 57
7. The status of null elements: Subjects in Sranan and Papiamentu 58
8. Partially overlapping categories: predicate adjectives in Creoles 59
9. Differences in lexical richness 60
10. Evidence from language contact 61
   10.1. Otomanguean-Spanish language contact 62
   10.2. Borrowing of content words in Salishan languages 65
11. Concluding remarks 67

Appendix 67

References 68
# Chapter 3
Philosophical Analysis as Cognitive Psychology: The Case of Empty Concepts

GEORGES REY

## Abstract

1. Introduction 73
2. Misadventures of the Classical View 73
3. Terminological issues 75
   - 3.1. Existential vs. (purely) intentional usage 75
   - 3.2. Concepts as between representations and referents 77
4. The inadequacies of Externalism 77
5. The need for internal roles 82
   - 5.1. The Quinean challenge 82
   - 5.2. The analytic data 83
   - 5.3. Rivals to the Analytic Explanation 85
6. Conclusion 87

## References

# Chapter 4
Categories and Cognitive Anthropology

JAMES S. BOSTER

## Abstract

1. Introduction 93
2. Cognition and culture, universalism and relativism 93
3. Paradigms and taxonomies 94
4. Kinship terminologies 101
5. Color classification 105
6. Ethnobiology 109
7. Towards a science of the stimulus 114

## References

# Chapter 5
Categorization in Neuroscience: Brain Response to Objects and Events

CATHERINE HANSON and STEPHEN JOSÉ HANSON

## Abstract

1. Introduction 120
2. Representing object categories in the brain
   - 2.1. Category-specific representation 122
   - 2.2. Feature-specific representation 123
   - 2.3. Process-specific representation 125
   - 2.4. Summary 125
3. Acquiring category knowledge
   - 3.1. Summary 130
4. Categorizing actions and events 130
   4.1. The nature of event knowledge 130
   4.2. When categorization of action fails 132
   4.3. The perception of events 133
   4.4. Summary 136
5. Conclusion 136
References 137

Chapter 6
Categorization in Cognitive Computer Science
JOHN F. SOWA
Abstract 141
1. Computation in cognitive science 142
2. The great categorization debates 144
3. From local features to global structures 148
4. Categorization and reasoning 153
5. Levels of cognition 158
References 161

PART 2 – SEMANTIC CATEGORIES

Chapter 7
Semantic Categorization
BRENDAN S. GILLON
Abstract 167
1. Introduction 168
2. The notional approach to lexical categories 169
3. The notional approach to lexical subcategories 170
4. Structural approach to semantic categories 172
5. Coordinators and subordinators 175
6. English nouns 181
7. Conclusion 184
Acknowledgments 184
References 184

Chapter 8
Emotion Categories across Languages
JAMES S. BOSTER
Abstract 188
1. Introduction 189
2. Methods of assessing cultural emotion systems 191
   2.1. The Method of Translation 191
   2.2. The Method of Mapping 198
Chapter 12
Definitions in Categorization and Similarity Judgments
SERGE LAROCHELLE, DENIS COUSINEAU and ANNIE ARCHAMBAULT
Abstract 278
1. Introduction 279
2. Importance rating and property selection 284
   2.1. Method 284
   2.2. Results 286
3. Categorization judgments 288
   3.1. Method 288
   3.2. Results 290
   3.3. Discussion 296
4. Similarity judgments 297
   4.1. Method 297
   4.2. Results 298
5. General discussion 300
References 302

Chapter 13
Why (Most) Concepts aren’t Categories
RUTH GARRETT MILLIKAN
Abstract 305
1. Introduction 306
2. Species are not categories 306
3. Three kinds of (Aristotelian) “substances” 307
   3.1. Historical kinds 307
   3.2. Eternal kinds 308
   3.3. Individuals 309
4. Concepts of individuals 310
5. Concepts of substances more generally 311
6. Substances encountered through language 312
References 315
PART 3 – SYNTACTIC CATEGORIES

Chapter 14
Lexical, Functional, Crossover, and Multifunctional Categories
LISA DEMENA TRAVIS

Abstract 320
1. Introduction 321
2. Categories as feature bundles 321
   2.1. The system 321
   2.2. Natural classes 322
   2.3. Unnatural classes 324
3. Categories and phrase structure 325
   3.1. Lexical and functional categories 326
   3.2. Articulation of functional categories 327
   3.3. Articulation below N and V 329
   3.4. Crossover and multifunctionality 330
4. Where do categorial distinctions reside? 336
5. Conclusions 344
References 345

Chapter 15
Isolating-Monocategorial-Associational Language
DAVID GIL

Abstract 348
1. Introduction 349
2. What IMA Language is Like 349
   2.1. Isolating 350
   2.2. Monocategorial 350
   2.3. Associational 351
3. Where IMA Language Is Found 354
   3.1. Semiotics 354
   3.2. Phylogeny 356
   3.3. Ontogeny 358
4. Typology 359
   4.1. Riau Indonesian: overview 360
   4.2. Riau Indonesian: analysis 364
   4.3. Riau Indonesian: A Relative IMA Language 374
5. Cognition 375
Acknowledgments 377
References 377
Chapter 16
Categories in Quebec Sign Language: Reflections on Categorization across Modalities
DENIS BOUCHARD, COLETTE DUBUISSON and ANNE-MARIE PARISOT
Abstract 381
1. The categories of lexical items 382
2. Traditional categorization applied to LSQ 384
   2.1. Nouns and verbs 387
   2.2. Pronouns and definite determiners 387
3. Pronouns in oral languages and in sign languages 388
   3.1. The effects of perceptual substances on linguistic forms 389
   3.2. Explaining the different properties 390
4. Consequences for linguistic categorization and universals 396
References 398

Chapter 17
Syntactic Categories in Signed versus Spoken Languages
DIANE LILLO-MARTIN
Abstract 402
1. Introduction 403
2. Lexical categories 403
3. Grammatical structures 405
   3.1. Subordination 405
   3.2. “Spatial syntax” 406
   3.3. Word order 413
4. Conclusion 417
Acknowledgments 418
Appendix. Notational conventions 418
References 419

Chapter 18
On Syntactic Categories
MARK C. BAKER 423

PART 4 – ACQUISITION OF CATEGORIES

Chapter 19
The Acquisition of Grammatical Categories: the State of the Art
MARIE LABELLE
Abstract 433
1. Grammatical categories 434
Table of Contents

2. Two-word utterances and their analysis 435
3. A semantic approach to grammatical categorization:
   Semantic bootstrapping 436
4. Distributional learning 439
   4.1. Word order 439
   4.2. Inflection and inflectional class 441
   4.3. Function words 445
   4.4. Word classes 447
   4.5. Other cues to grammatical category learning 449
5. Models of distributional learning 449
6. Constraining the search space 450
7. Conclusion 451
References 452

Chapter 20
Semantic Categories in Acquisition
EVE VIVIENNE CLARK
Abstract 459
1. Introduction 460
2. Space 461
3. Shape 465
4. Adding common ground 466
5. Conceptual domains and lexical options 467
6. Adding meaning in the course of conversation 472
7. Universals in mapping? 473
8. Conclusion 476
References 477

Chapter 21
Early Syntactic Categories in Infants' Language
RUSHEN SHI
Abstract 481
1. Introduction 482
2. The acquisition of grammatical categories and the earliest binary distinction of function words and content words 482
3. Input speech and the categorization of function words and content words 486
4. Function words and language acquisition 488
5. Conclusions 492
Acknowledgment 492
References 493
### Chapter 22

Acquiring Auditory and Phonetic Categories  
MARTIJN GOUBEEK, ROEL SMITS, ANNE CUTLER  
and DANIEL SWINGLEY

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>497</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>498</td>
</tr>
<tr>
<td>2. Testing category learning</td>
<td>500</td>
</tr>
<tr>
<td>3. Learning of nonspeech categories</td>
<td>502</td>
</tr>
<tr>
<td>4. Learning of speech categories</td>
<td>506</td>
</tr>
<tr>
<td>5. Conclusion</td>
<td>510</td>
</tr>
<tr>
<td>References</td>
<td>511</td>
</tr>
</tbody>
</table>

### Chapter 23

Syntactic Categories in Second Language Acquisition  
LYDIA WHITE

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>515</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>516</td>
</tr>
<tr>
<td>2. Lexical and functional categories</td>
<td>516</td>
</tr>
<tr>
<td>3. Lexical categories in L2 acquisition</td>
<td>517</td>
</tr>
<tr>
<td>4. Functional categories in acquisition: Issues of evidence</td>
<td>519</td>
</tr>
<tr>
<td>5. Functional categories in the L2 initial state and in L2 development</td>
<td>522</td>
</tr>
<tr>
<td>5.1. Morphology-before-syntax</td>
<td>522</td>
</tr>
<tr>
<td>5.2. Syntax-before-morphology</td>
<td>523</td>
</tr>
<tr>
<td>6. Acquiring versus losing categories and features</td>
<td>524</td>
</tr>
<tr>
<td>7. Discussion</td>
<td>529</td>
</tr>
<tr>
<td>References</td>
<td>530</td>
</tr>
</tbody>
</table>

### Chapter 24

The Development of Categories in the Linguistic and Nonlinguistic Domains: the Same or Different?  
DIANE POULIN-DUBOIS

### PART 5 – NEUROSCIENCE OF CATEGORIZATION AND CATEGORY LEARNING

### Chapter 25

Multiple Systems of Perceptual Category Learning: Theory and Cognitive Tests  
F. GREGORY ASHBY and VIVIAN V. VALENTIN

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>548</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>549</td>
</tr>
<tr>
<td>2. Two Category-Learning Tasks</td>
<td>550</td>
</tr>
<tr>
<td>3. COVIS</td>
<td>550</td>
</tr>
<tr>
<td>4. The COVIS explicit system</td>
<td>552</td>
</tr>
<tr>
<td>4.1. Switching attention in the explicit system</td>
<td>554</td>
</tr>
<tr>
<td>4.2. Long-term storage of explicit category knowledge</td>
<td>556</td>
</tr>
</tbody>
</table>
Chapter 26
The Neuropsychology of Perceptual Category Learning
W. TODD MADDOX and J. VINCENT FILOTEO
Abstract
1. Introduction
2. Competition between verbal and implicit systems (COVIS)
3. Testing a priori Predictions of COVIS
4. Perceptual category learning in neurological patients
   4.1. Nonlinear information-integration category learning in amnesia
   4.2. Nonlinear information-integration category learning in striatal-damaged patients
   4.3. Rule-based category learning in PD
   4.4. Further study of information-integration category learning in PD
5. General discussion
References

Chapter 27
Neural Regions Associated with Categorical Speech Perception and Production
SUSAN M. RAVIZZA
Abstract
1. Introduction
2. Evidence for categorical speech processing
3. Prefrontal regions and motor speech categories
4. Temporal–parietal regions and acoustic speech categories
5. Cerebellar contributions to categorical production and perception
6. Concluding remarks
References

PART 6 – CATEGORIES IN PERCEPTION AND INFERENCE

Chapter 28
Situated Conceptualization
LAWRENCE W. BARSALOU
Abstract
Chapter 29
Perceptual and Semantic Reorganization during Category Learning
ROBERT L. GOLDSTONE, BRIAN J. ROGOSKY, RACHEL PEVTZOW
and MARK BLAIR
Abstract 652
1. Introduction 653
2. Concept learning and perception 654
  2.1. Object segmentation 654
  2.2. Experiment 1 655
  2.3. Experiment 2 657
  2.4. Conclusions on perceptual reorganization 662
3. Semantic reorganization during category learning 664
  3.1. Integral versus separable dimensions 664
  3.2. Experiment 3 665
  3.3. Experiment 4 669
4. Conclusions on semantic reorganization 671
5. Integrating perceptual and semantic reorganization 671
  5.1. Characterizing psychological features 672
  5.2. Characterizing featural change 673
  5.3. Prospects for synthesizing perceptual and semantic reorganization 675
Acknowledgments 676
References 676

Chapter 30
The Return of Concept Empiricism
JESSE J. PRINZ
Abstract 679
# Table of Contents

1. Introduction 680
2. Concept empiricism 680
   2.1. Representing and doing: Two faces of concepts 680
   2.2. Variable mechanisms 684
   2.3. Perceptual vehicles 686
   2.4. Innateness 688
   2.5. Summary 691
3. The abstract ideas objections 692
References 694

## PART 7 – GROUNDING, RECOGNITION, AND REASONING IN CATEGORIZATION

### Chapter 31
Categorization, Reasoning, and Memory from a Neo-logical Point of View
SERGE ROBERT
Abstract 700
1. Introduction 701
2. Order, Information, and Categories 701
3. Inferences, Arguments, and Information 701
4. Logical and Creative Arguments 703
5. Types of Creative Arguments 703
6. Two Rather Incompatible Views in Cognitive Science 705
7. Experiments on Memory and Logical Competence 706
8. Logical Weakness and Classification of Errors 709
9. A New Theory of Logical Error and Logical Competence 710
10. The Cognitive Functions of Logic 711
11. Corrective Inferences and Cognitive Progress 713
12. The Fundamental Cognitive Function of Logical Reasoning 715
13. The Mind as a Dynamic System: Inference and Memory 715
14. General Conclusions: Categorization, Reasoning, and Memory 716
References 717

### Chapter 32
Approaches to Grounding Symbols in Perceptual and Sensorimotor Categories
ANGELO CANGELOSI
Abstract 719
1. Cognitive symbol grounding 720
   1.1. The symbol grounding problem 720
   1.2. Grounding symbols in cognition 721
2. Linking vision and language: connectionist approaches to category learning and symbol grounding 724
   2.1. Connectionist modeling of category learning and naming 724
   2.2. Connectionist modeling of symbol grounding transfer 727
Chapter 33
Embodied Categorization
PIERRE POIRIER, BENOIT HARDY-VALLÉE and
JEAN-FRÉDÉRIC DEPASQUALE

Abstract

1. Introduction: Embodied categorization
2. Purely reactive categorizers
   2.1. The perceptual aliasing problem [Whitehead and Ballard (1991)]
   2.2. Type I versus II problems [Clark and Thornton (1997)]
3. Reactive categorizers that learn
4. Representing categorizers
5. Emulating and simulating categorizers
   5.1. Emulating categorizers
   5.2. Simulating categorizers
6. Analogizing categorizers
7. Linguistic categorizers
8. Conclusion
References

Chapter 34
Categorization of Objects, Scenes, and Faces through Time
ÉRIC McCABE, CAROLINE BLAIS and FRÉDÉRIC GOSELIN

Abstract

1. A model of categorization
2. Basic-level literature
3. Discrete processing cycles
   3.1. A Bubbles primer
   3.2. Fossilized discrete processing cycles
   3.3. What can temporal bubbles reveal about a SLIP categorizer?
4. The need for flexibility and a paradox
   4.1. Limited processing capacity
   4.2. The need for flexibility
   4.3. Back to the paradox
5. Categorization as an iterative process
   5.1. Compulsory feedforward processing sweeps
   5.2. Flexible iterative processing sweeps
6. General discussion
References
Chapter 35
Adaptive Categorization and Neural Networks
ROBERT PROULX and SÉBASTIEN HÉLIE

Abstract

1. The problem of divergence
2. The solution: dual Hebbian/anti-Hebbian learning
   2.1. Stabilization
   2.2. Oscillation
   2.3. Linearity
   2.4. Additional properties of the learning rule
3. The Eidos model
4. The letter classification task
   4.1. Methodology
   4.2. Results
5. The problem of convergence
6. The solution: Unlearning
7. The letter-classification task revisited
   7.1. Methodology
   7.2. Results
8. Current trends: Elimination of spurious attractors
9. Conclusion
References

Chapter 36
A Grounded Mind in a Robotic Body
STEVAN HARNAD

PART 8 – MACHINE CATEGORY LEARNING

Chapter 37
Concept Learning and Nonmonotonic Reasoning
PETER GÅRDENFORS

Abstract

1. The role of concepts
2. Three kinds of cognitive representations
3. Learning in symbolic systems
4. Learning in connectionist systems
5. Conceptual spaces as a representational framework
6. The origin of quality dimensions
7. Properties and concepts
8. Prototypes and conceptual spaces
9. Learning in conceptual spaces
10. The role of similarity in learning
11. Nonmonotonic aspects of concepts

References
# Table of Contents

7.1. Main steps 863
7.2. Descriptive SDA in SODAS 864
8. Overview of SODAS 865
  8.1. Some advantages of the use of concepts modeled by symbolic objects 865
  8.2. Overview of SODAS software 865
9. Final remarks 865
References 866

**Chapter 39**
Category Formation in Self-organizing Embodied Agents
STEFANO NOLFI

Abstract 869
1. Introduction 870
2. The method 870
3. Categories emerging from the interaction between
   the agent and the environment 871
   3.1. Finding and remaining in favorable environmental areas 871
   3.2. Discriminating objects with different shapes on the basis of tactile information 873
   3.3. Behavior emerging from the dynamic interaction between the agent and its environment 876
4. Action-mediated sensory states 878
   4.1. Discriminating larger and smaller cylindrical objects 879
   4.2. Navigating toward a target area of the environment 881
5. Integrating sensorimotor information over time and the emergence of
   complex internal categories 883
   5.1. The self-localization problem 884
6. Conclusions 887
Acknowledgments 888
References 888

**Chapter 40**
An Information-based Discussion of Vagueness:
Six Scenarios Leading to Vagueness
DIDIER DUBOIS and HENRI PRADE,
FRANCESC ESTEVA and LLUIS GODO

Abstract 892
1. Introduction 893
2. The information framework 894
3. Classical vs. gradual properties 895
   3.1. Graduality and partial preordering 895
   3.2. Membership functions as total preorders 896
   3.3. Fuzzy sets and similarity to prototypes 897
   3.4. Set-theoretic operations 897
3. Standardization

4. Grammatical tagging
   4.1. Why expert rules at the tagging stage?
   4.2. A tagging language
   4.3. Our approach to grammatical tagging
   4.4. Automatic learning of new tagging rules

5. Terminology

6. Concept recognition in texts
   6.1. Polysemy
   6.2. General versus local collocations
   6.3. Terms and collocations
   6.4. ACT as a friendly interface helping the expert
   6.5. ACT as an inductive program
   6.6. Validation

7. Conclusion

Acknowledgments

References
Chapter 44
Graph Matching, System Design and Knowledge Modeling
GUY W. MINEAU
Abstract
1. Introduction
2. Knowledge represented as graph structures
3. Learning heuristic knowledge
4. Viability conditions
5. The complexity of learning
6. Categorization of knowledge in layers
7. Conclusion
References

PART 10 – THE NATURALIZATION OF CATEGORIES

Chapter 45
Nominalism and the Theory of Concepts
CLAUDE PANACCIO
Abstract
1. Nominalism
2. Ockham’s cleaver
3. Motivations
4. Nominalistic constraints for the theory of concepts
   4.1. Represented things as singular
   4.2. Representations as singular
References

Chapter 46
Why do We Think Racially?
EDOUARD MACHERY, LUC FAUCHER
Abstract
1. Introduction
2. Is racialism a mere social construct?
   2.1. Racial skepticism
   2.2. Races are interactive kinds
   2.3. Races are transient kinds
   2.4. Merits and problems
3. Is racialism a by-product of a human kind module?
Chapter 47
Neurosemantics and Categories
CHRIS ELIASMITH
Abstract
1. Introduction
  1.1. Why “neuro”?
  1.2. The explanandum
2. Mental representations as neural codes
  2.1. Representations
  2.2. Transformation
  2.3. A representational hierarchy
3. The meaning of neural representations: Neurosemantics
  3.1. The representation relation
  3.2. A neurosemantic theory
  3.3. Discussion
4. Misrepresentation
5. Conclusion
References

Chapter 48
Conceptual Analysis and Philosophical Naturalism
ELISABETTA LALUMERA
Abstract
1. Introduction
2. What is intuitive about conceptual analysis?
3. Cognitive privileges, metaphysical privileges, and the Transparency Thesis
4. Against privileges
References
<table>
<thead>
<tr>
<th>Chapter 49</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Crisis! What Crisis?</td>
<td>1073</td>
</tr>
<tr>
<td>PIERRE POIRIER</td>
<td></td>
</tr>
</tbody>
</table>

Index

1081
LIST OF CONTRIBUTORS

**Archambault, Annie**
Département de Psychologie, C.P. 6128, Succ. Centre-ville, Montreal, QC, H3C 3J7 Canada.

**Ashby, F. Gregory**
Department of Psychology, University of California, Santa Barbara, CA 93106, USA.

**Baker, Mark C.**
Department of Linguistics, Rutgers University, 18 Seminary Place, New Brunswick, NJ 08901, USA.

**Barsalou, Lawrence W.**
Department of Psychology, Emory University, Atlanta, GA, 30329, USA.

**Biskri, Ismail**
Département de mathématique et d’informatique, Université du Québec à Trois-Rivières, 3351 Boul. des Forges, Trois-Rivières, G9A 5H7 Canada.

**Blair, Mark**
Psychology building, 1101 E 10th Street, Indiana University, Bloomington, IN 47405-7007, USA.

**Blais, Caroline**
Département de Psychologie, Université de Montréal, C.P. 6182, Succ. Centre-ville, Montreal, QC, H3C 3J7 Canada.

**Boster, James S.**
Department of Anthropology, U-2176, 354 Mansfield Road, University of Connecticut, Storrs, CT 06269-2176, USA.

**Bouchard, Denis**
Département de Linguistique, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.
Cangelosi, Angelo
Adaptive Behaviour and Cognition Research Group, School of Computing, Communications and Electronics, University of Plymouth, Portland Square A316, Plymouth PL4 8AA, UK.

Clark, Eve V.
Department of Linguistics, Margaret Jacks Hall (Bldg 460), Stanford University, Stanford, CA 94305-2150, USA.

Cohen, Henri
Cognitive Neuroscience Center, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Cook, Richard S.
University of California at Berkeley, 1203 Dwinelle Hall, Berkeley, CA 94720-2650, USA.

Cousineau, Denis
Université de Montréal, C.P. 6128, Succ. Centre-ville, Montreal, QC, H3C 3J7 Canada.

Cutler, Anne
Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands.

De Pasquale, Jean-Frédéric

Diday, Edwin
Paris IX Dauphine University, Place du Maréchal de Lattre de Tassigny, Bureau B633, 75775 Paris cedex 16, France.

Dubois, Didier
CNRS, Porte 308, IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse, cedex 4, France.

Dubuisson, Colette
Département de Linguistique, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Eliasmith, Chris
Department of Philosophy, University of Waterloo, Waterloo, ON, N2L 3G1 Canada.
Esteva, Francesc
III - Institut d’Investigació en Intel.ligència Artificial, CSIC – Spanish Scientific Research Council, Campus Universitat Autonoma de Barcelona, 08193 Bellaterra, Catalonia, Spain.

Faucher, Luc
Département de philosophie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Filoteo, J. Vincent
UCSD/VA, 3350 La Jolla Village Drive, San Diego, CA 92161-116A, USA.

Forest, Dominic
Département de Philosophie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Gärdenfors, Peter

Gil, David
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany.

Gillon, Brendan S.
Department of Linguistics, McGill University, 1085 Doctor Penfield Avenue, Montreal, QC, H3A 1A7, Canada.

Godo Lacasa, Lluís
Institut d’Investigació en Intel.ligència Artificial (IIIA), Consejo Superior de Investigaciones Científicas (CSIC), 08193 Bellaterra, Spain.

Goldstone, Robert R.
Psychology building, 1101 E 10th Street, Indiana University, Bloomington, IN 47405-7007, USA.

Gosselin, Frederic
Département de Psychologie, Université de Montréal, C.P. 6182, Succ. Centre-ville, Montreal, QC H3C 3J7 Canada.

Goudbeek, Martijn
Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands.

Hanson, Catherine
Psychology Department, Rutgers University, 101 Waren St., Newark, NJ 07102, USA.
Hanson, Stephen Jose  
Psychology Department, Rutgers University, 101 Waren St., Newark, NJ 07102, USA.

Hardy-Vallée, Benoît  
Institut Jean-Nicod, UMR 8129, 1bis, Avenue de Lowendal, F-75007 Paris, France.

Harnad, Stevan  
Cognitive Neuroscience Center, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville Montreal, QC, H3C 3P8 Canada.

Hélie, Sébastien  
Laboratoire d’Études en Intelligence Naturelle et Artificielle, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Kay, Paul  
International Computer Science Institute, Berkeley, ICSI, 1947 Center St, Suite 600 Berkley, CA 94704-1198, USA.

Kodratoff, Yves  
CNRS, LRI, Université Paris-Sud XI, 91405 Orsay, France. Tel.: 33-1 69156904; Fax: 33-1 69156586; E-mail: yk@lri.fr

Labelle, Marie  
Département de linguistique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Lalumera, Elisabetta  
Dipartimento di Discipline della Comunicazione, Università degli Studi di Bologna, Via A. Gardino 23, 40131, Bologna, Italy.

Larochelle, Serge  
Département de Psychologie, Université de Montréal Université de Montréal, C.P. 6128, Succ. Centre-ville, Montreal, QC, H3C 3J7 Canada.

Lefebvre, Claire  
Département de linguistique et didactique des langues, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Lillo-Martin, Diane  
Department of Linguistics, University of Connecticut, 337 Mansfield Road, Unit 1145, Storrs, CT 06269-1145, USA.

Machery, Edouard  
Department of History and Philosophy of Science, University of Pittsburgh, CL 1017, Pittsburgh, PA 15260, USA.
Maddox, W. Todd
Department of Psychology and Institute for Neuroscience, 1 University Station A 8000, University of Texas, Austin, TX, 78712, USA.

McCabe, Éric
Département de Psychologie, Université de Montréal, C.P. 6182, Succ. Centre-ville, Montreal, QC, H3C 3J7 Canada.

Meunier, Jean-Guy
Département de Philosophie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Millikan, Ruth* Gerrett
Department of Philosophy, U-2054, University of Connecticut, Storrs, CT 06269-2054, Office: 202 Manchester Hall, USA.

Mineau, Guy
Faculty of Sciences and Engineering, Department of Computer Science and Software Engineering, Université Laval, Québec, QC, G1K 7P4 Canada.

Muysken, Pieter
Linguistics, Radboud University, Nijmegen. Postbus 9103, Nijmegen, The Netherlands.

Napoli, Amedeo
LORIA, LORIA, B.P. 239, 54506 Vandœuvre les Nancy, France.

Nolfi, Stefano
Institute of Cognitive Sciences and Technologies, National Research Council (CNR), Viale Marx 15 00137 Rome, Italy.

Panaccio, Claude
Département de Philosophie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Papafragou, Anna
108 Wolf Hall, University of Delaware, Newark, DE 19716, USA.

Parisot, Anne-Marie
Département de Linguistique, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Pevtzow, Rachel
Psychology building, 1101 E 10th Street, Indiana University, Bloomington, IN 47405-7007, USA.

* Important: Pending upon Elsevier’s permission to reproduce
Poirier, Pierre
Département de Philosophie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Poulin-Dubois, Diane
Psychology Department, Concordia University, 7141 Sherbrooke Street West, Montreal, QC, H4B 1R6 Canada.

Prade, Henri
CNRS, porte 307, IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062, Toulouse, cedex 4, France.

Prinz, Jesse J.
Department of Philosophy, CB#3125, Caldwell Hall, University of North Carolina, Chapel Hill, NC 27599, USA.

Proulx, Robert
Faculté des Sciences Humaines, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3C 3P8 Canada.

Ravizza, Susan M.
Department of Psychology, UC Davis Imaging Research Center, 4701 X Street, Sacramento, CA 95817, USA.

Regier, Terry
Department of Psychology, University of Chicago, 5848 S. University Avenue, Office: Green 414, Chicago, IL 60637, USA.

Rey, Georges
Department of Philosophy, University of Maryland, College Park, MD 20742, USA.

Robert, Serge
Department of Philosophy, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC H3C 3P8 Canada.

Rogosky, Brian J.
Psychology building, 1101 E 10th Street, Indiana University, Bloomington, IN 47405-7007, USA.

Shi, Rushen
Département de Psychologie, Université du Québec à Montréal, C.P. 8888 Succ. Centre-ville, Montreal, QC, H3c 3P8 Canada.
Smits, Roel
Max Planck Institute for Psycholinguistics, Wundtlaan 1, P.O. Box 310, 6500 AH Nijmegen, The Netherlands.

Sowa, John F.
VivoMind LLC, 21 Palmer Avenue, Croton-on-Hudson, NY 10520, USA. Tel.: 914-271-5557; E-mail: sowa@bestweb.net

Swingley, Daniel
Department of Psychology, University of Pennsylvania, 3401 Walnut Street 302 C, Philadelphia, PA 19104, USA.

Thagard, Paul
Philosophy Department, University of Waterloo, Waterloo, ON, N2L 3G1 Canada. Tel.: 519-888-4567, ext 3594; Fax: 519-746-3097; E-mail: pthagard@uwaterloo.ca

Toombs, Ethan
Philosophy Department, University of Waterloo, Waterloo, ON, N2L 3G1 Canada.

Travis, Lisa deMena
McGill University, 1085 Dr. Penfield Avenue, Department of Linguistics, Montreal, QC H3A 1A7 Canada.

Valentin, Vivian V.
Department of Psychology, University of California Santa Barbara, CA 93106, USA.

White, Lydia
Department of Linguistics, McGill University, 1085 Dr. Penfield Avenue, Montreal, QC, H3A 1A7 Canada.
This page intentionally left blank
Contents

1. Introduction 2
2. Organization of the book 2
3. Major common themes 7
   3.1. The notions of category and categorization 7
   3.2. The nature of categories: Discrete, vague, or other 9
   3.3. Are there modality effects on categories? 10
   3.4. Are there universal categories? Are there innate categories? 11
4. Bridging the category divide 13
References 15
1. Introduction

Categorization is the mental operation by which the brain classifies objects and events. This operation is the basis for the construction of our knowledge of the world. It is the most basic phenomenon of cognition, and consequently the most fundamental problem of cognitive science. Cognitive science is concerned with the kinds of knowledge that underlie human cognition, the details of human cognitive processes, and the computational modeling of these processes.

This book presents the study of categories and the process of categorization as viewed through the lens of the founding disciplines of the cognitive sciences: cognitive anthropology, cognitive computer science, linguistics, neuroscience, philosophy, and psychology. The study of categorization has long been at the core of each of these disciplines.

The literature on categorization reveals that there is a plethora of definitions, theories, models, and methods to apprehend this central object of study. The contributions in this handbook reflect this diversity. For example, the notion of category is not uniform across these contributions and there are multiple definitions of the notion of concept. Furthermore, the study of category and of categorization is approached differently within each discipline. For some authors, the categories themselves constitute the object of study, whereas for others, it is the process of categorization, and for others still, it is the technical manipulation of large chunks of information. Finally, yet another contrast has to do with the biological versus artificial nature of agents or categorizers.

Recently, since cognitive science came on the scene, there has been a concerted effort to establish connections between disciplines. As a result, our understanding of human cognition has been profoundly altered. This book constitutes a major effort to bring the various disciplines together, for the first time, around a single theme: categorization. We hope that this collective work will result in the cross-fertilization of methods and ideas, and that it will contribute significantly to our understanding of categorization in particular, and of human cognition in general.

In the sections that follow, we present the contents of the book. We then summarize the major themes and issues that are raised, and we point out some of the controversies and similarities that emerge between the authors and disciplines represented here.

2. Organization of the book

The book is organized in 10 parts. The first one, Categorization in Cognitive Science, contains six chapters introducing the reader to the notion of category/categorization from the point of view of the six disciplines involved. From the perspective of psychology, Harnad considers that categorization is the most fundamental cognitive process and concludes that to cognize is to categorize. From the point of view of linguistics, Muysken shows that grammatical categories are not unitary notions, but emerge at the interface of different components of our human cognitive and communicative capacities. Rey discusses how philosophers approach the notion of concept and
argues for the importance of empty concepts. Boster provides an overview of the representations of category structure by cognitive anthropologists in kinship terminologies, color classification, and ethnobiology. From a neuroscientific perspective, Hanson and Hanson look at how the brain represents category knowledge and at what information is acquired during category learning. Finally, Sowa surveys theories of categorization and reasoning in cognitive science that have been implemented and tested in computer systems. He also shows that, while most of the ideas emerged before the advent of computers, this technology did provide the tools for developing these ideas in a way that had previously been impossible.

The other nine parts are organized by theme rather than by discipline. These themes are semantic categories; syntactic categories; acquisition of categories; neuroscience of categorization and category learning; categories in perception and inference; grounding, recognition, and reasoning in categorization; machine category learning; data mining for categories and ontologies; and finally, the naturalization of categories. In general, the first chapter of each part constitutes an introduction to the theme.

In the first chapter on semantic categories in Part 2, Gillon addresses the question of what a semantic category is. After presenting the two views from which linguists have addressed this question – the structural and the notional – he argues that only the structural view has provided sound answers to linguistic questions. The next chapter, by Boster, discusses the degree to which emotions are biologically endowed or culturally constructed. It explores the nature of the similarities and differences among cultural emotion systems, and compares methods of carrying out that exploration. Cook, Kay, and Regier present the history of the World Color Survey, and show how this database has been used to test the universality of color naming across languages. On the basis of the observation that the concept of [ATOM] has undergone numerous changes throughout the history of chemistry, Thagard and Toombs use this concept to discuss the relationship between categorization and conceptual change. Papafragou considers the question of whether crosslinguistic differences in marking the count/mass distinction in syntax affect the nonlinguistic individuation criteria used by speakers of different languages. Larochelle, Cousineau, and Archambault examine the role of definitions in categorization and similarity judgments. They show that, while definitions and characteristic descriptions are treated alike when making similarity judgments, definitions appear to be treated in a more unitary fashion when making categorization judgments. Finally, Millikan takes up the challenge of explaining why most concepts are not categories.

Part 3 on syntactic categories begins with Travis’s introduction to types of syntactic categories: lexical or major categories, functional or minor categories, crossover categories (e.g., gerunds), and multifunctional categories. While her main goal is to provide some background on the issues raised by the study of syntactic categories, she defends the need for categorial information in lexical entries. Although most languages display a number of syntactic categories, there appear to be languages that are almost deprived of them. Based on the study of Riau Indonesian, Gil argues that languages of the latter type share three basic properties. They are morphologically isolating and present no
word-internal morphological structure, they are monocategorial, and they are semantically associational. From variation in syntactic categories between spoken languages, we turn to differences between languages that may be due to differences in modality. Based on data from Quebec Sign Language, Bouchard, Dubuisson, and Parisot show that categorization is modality-dependent, and that therefore, some categories are not the same in signed and in spoken languages. Lillo-Martin further explores the similarities and differences between spoken and signed languages. She shows that, while there are profound differences between spoken and signed languages (e.g., the use of spatial information to convey information about space), in some cases, morphosyntactic properties that at first glance seemed to show effects of the linguistic modality, are in fact underlyingly similar across the modes. Both chapters explore the consequences of the modality effect on syntactic categories for the notion of linguistic universals. This part of the book ends with a discussion by Baker, who takes a stand in favor of the universality of grammatical categories both across languages and across modalities (oral and sign).

The fourth part, dedicated to the acquisition of categories, is introduced by Labelle’s presentation of the state of the art. She reviews recent research on the acquisition of grammatical categories, focusing on three aspects of the problem – parts of speech, inflection, and subcategories of words – and she reviews the various proposals suggested to account for the fact that children appear to master the complexity of syntactic categories quite early in their development of language. Considering that both universal and language-specific meanings play a role in children’s acquisition of semantic categories, Clark weighs the respective contributions of universal conceptual categories and of the conventions of the language community in the acquisition of these categories. Shi considers early syntactic categories in infants, proposing that infants can derive the distinction between content words and function words on the basis of a constellation of acoustic/phonetic and phonological cues in the input. Goudbeek, Smits, Swingley, and Cutler consider the acquisition of auditory and phonetic categories. This learning is viewed as the formation of categories in a multidimensional psychophysical space. White’s chapter bears on second language acquisition, and more specifically on how syntactic categories (lexical and functional) are acquired and represented in the interlanguage grammars of second language (L2) speakers. Several issues are addressed, including whether L2 learners simply adopt the syntactic categories that are represented in their L1. In her discussion of the chapters on acquisition, Poulin-Dubois shows how the issues that are relevant to the development of categories in the language domain both overlap with, and differ from, those that are raised in the nonlinguistic domain.

Part 5 focuses on the neuroscience of categorization and category learning. With data from animal, lesion, neuropsychological, and computer-modeling studies, neuroscience has recently witnessed a wealth of results that are homing in on the neural mechanisms and structures that mediate category learning. In the first chapter in this part, Ashby and Valentin present COVIS, a well-developed theoretical model of perceptual category learning, in which two learning systems are set in opposition. The frontal-based explicit system depends on working memory and executive attention, and is mediated primarily by the anterior cingulate, the prefrontal cortex, and the head of the caudate; the basal
ganglia-mediated system uses procedural learning and requires a dopamine reward signal. Their colleagues Maddox and Filoteo investigate how the predictions made by COVIS accord with data from neurological patients. In the last chapter, Ravizza shows that the ease with which we produce and comprehend speech belies the large number of neural areas supporting these skills. She presents neuropsychological and neuroimaging evidence corroborating the idea that speech processes are categorical. These three contributions highlight the role of the specific brain structures involved in categorization.

Part 6 addresses questions related to categories in perception and inference. Our conceptual system is made up of category representations and is central to memory, language, and thought. Drawing on psychology and cognitive neuroscience, Barsalou makes the case that a given concept can produce different situated conceptualizations, each tailored to different instances in different settings. In a somewhat complementary vein, Prinz takes sides in the debate between rationalists and empiricists, and defends the view that conceptual representations are perceptually based. In the last chapter, we learn that our representation of objects not only influences, but is influenced by the concepts that we learn. Goldstone, Rogosky, Petzow, and Blair present evidence that demonstrates how categorization experience alters the descriptions of objects. In category learning, we create the elements of categorized objects’ descriptions and, at the same time, associate those elements with categories.

Grounding, recognition, and reasoning in categorization are the topics covered in Part 7. Much of our knowledge is the result of categorization, and reasoning plays an important role in this process. In the first chapter, Robert investigates the relations between categorization and reasoning, with some discussion of memory, to gain a better understanding of the mechanisms involved in our capacity to build categories. The issue of intrinsically linking the symbols used by a cognitive agent to their corresponding meanings has been called “the symbol grounding problem.” Cangelosi presents connectionist and embodied modeling approaches for the grounding of language in perception, cognition, and action. Poirier, Hardy-Vallée, and DePasquale further explore the embodied nature of categorization by studying the categorizing abilities of architecturally simple embodied agents, then turn to living systems, switching from AI and robotics to neuroscience and psychology. Their concern is to show what these living systems have in common with their artificial counterparts, in an effort to determine how embodiment influences categorization. In the next chapter, McCabe, Blais, and Gosselin address some aspects of categorization that have been neglected in the literature: categorization through time, limited processing capacities, and the paradox that this creates. Finally, connectionist networks of categorization rely on Hebbian learning to convert the stimulus space into a feedback subspace sufficient to categorize new stimuli. The Hebbian learning rule specifies by how much the weight of the connection between two units should be increased or decreased in proportion to the product of their activation. The rule builds on Hebb’s (1949) learning rule, postulating that the connections between two neurons might be strengthened if they fire simultaneously. Proulx and Hélie test a new learning/unlearning procedure applied to an existing model, and review a new approach that reduces the number of spurious attractors. This part of the book ends with a discussion by Harnad.
Part 8 addresses issues related to machine category learning. Gärdenfors argues that there are aspects of cognitive phenomena for which neither symbolic representation nor associationism seems to offer appropriate modeling tools. In the first chapter of this section, he advocates a third form of representing information, which employs geometric structures rather than symbols or associations, outlining how conceptual spaces can be used to model some of the fundamental aspects of how we learn and reason with concepts. Diday adopts a computer sciences-based approach to categories and concepts as they emerge from databases. He attempts to show how these notions from cognitive science can improve knowledge discovery in text and data mining and how, in return, symbolic data analysis can extract categories from concepts. Nolfi focuses on how categories might emerge, through artificial evolution, from the dynamic interaction between situated agents and their environment, and on the relation between categories and behavior. Finally, it seems that vagueness can be precisely defined: a concept is vague as soon as it partitions the universe of discourse. In their chapter, Dubois, Esteva, Godo, and Prade discuss and investigate a series of information scenarios concerning vagueness from an AI perspective by focusing on knowledge representation.

Part 9 deals with data mining for categories and ontologies. Extracting information from databases is like searching for gold in riverbeds. In the first chapter, Napoli introduces the knowledge discovery process, encompassing the requisite steps for efficient mining operations in very large databases. To solve the problem of concept recognition in texts, Kodratoff defines a new form of learning called extensional induction. The goal of these inductive procedures is not to define a concept but to help field experts recognize its presence in a text. Meunier, Forest, and Biskri also work closely with specialist readers, and show how computer-assisted reading and analysis can help them sift through the content, themes, and concepts of large textual corpora. Modeling complex systems can be likened to a well-honed art where the designer must first ask whether the modeling activities interact with the world with sufficient complexity. In the graph-matching problem, Mineau explores this question and illustrates the relationship between knowledge-based technology and the modeling required to produce such systems.

The role of philosophers in cognitive science is illustrated in the chapters in the last part of the book, dealing with the naturalization of categories. In his defense of nominalism – the claim that only singular objects and individuals exist – Panaccio explains how a nominalistic outlook can bear upon the way we think about concepts, making us aware of certain important problems and distinctions, and avoid widespread confusion among the things we are talking about when we talk about concepts. For Eliasmith, fundamental assumptions about semantics remain implicit in most discussions about categorization. He argues that categorization is a universal neural phenomenon, and attempts to outline a semantics for computational neuroscience where neural states are representations. In Machery and Faucher’s view, the time has come to bridge the gap between social constructionism and cognitive evolutionary theories of racial categorization. Drawing on several disciplines, they attempt to explain why, of all the ways of dividing the world, we find that some categories, but not others, are natural. In the last chapter, Lalumera contrasts two positions that a naturalistically minded philosopher
may take toward the intuition of special-status contents. She defends an alternative view according to which relations among concepts mirror the relations among the real-world properties they refer to. Poirier concludes this part of the book with an overview of the contribution of philosophers to cognitive science.

3. Major common themes

A number of common themes run through the chapters in this book. They include the notions of category and categorization, the nature of categories – whether discrete, vague, or other – whether there is a modality effect on categories, and finally, whether there are categories that are universal and categories that are innate. In this section, we summarize what our authors have to say about these themes.

3.1. The notions of category and categorization

The first question that one may ask is whether the notion of category is uniform across disciplines and authors. Looking back at the contents of Part 1, one may already surmise the answer to this question. Harnad focuses on cognitive states and processes, Muysken on grammatical categories, Rey on concepts and empty concepts, Boster on structural arrangements of categories, Hanson and Hanson on how the brain – and the damaged brain – represents category knowledge, and Sowa on artificial system categories. In the following paragraphs, we present the various ways that the notions of category and categorization are being used by the various authors in this book.

The linguists refer to phonetic, phonological, syntactic, semantic, and lexical categories. All these categories may be described in terms of feature bundles. For example, major lexical categories are defined by a combination of the major features [+/-N(oun)], [+/-V(erb)], yielding the four major lexical categories in (1) [Chomsky (1970)]:

\[(1) \begin{align*}
[+ N, - V]: & \text{Nouns} \\
[- N, + V]: & \text{Verbs} \\
[+ N, + V]: & \text{Adjectives} \\
[- N, - V]: & \text{Pre-/postpositions}
\end{align*}\]

Minor or functional categories are defined in terms of minor features such as [+/-Det(eminier)], [+/-T(ense)], [+/-P(iteral)], etc. Likewise, semantic categories may be defined by features such as [+/-human], [+/-abstract], and so on. Although there is debate concerning some cases (e.g., the status of prepositions as major- or minor-category lexical items), there is a general consensus among linguists as to what constitutes a grammatical category.

For cognitive anthropologists, the notion of category is not far removed from that of linguists. In their view, categories (e.g., color terms, kinship terms, etc.) reveal themselves through the lexicon, and therefore, the basic task of the investigator consists in exploring the cognitive organization of the lexicon for a given domain, with componential analysis
as an important tool. Several authors, whose studies are reported by Boster, make use of feature bundles in referring to folk categories.

In contrast to linguists and cognitive anthropologists, other contributors focus on concepts or the process of categorization as their object of study. As Rey discusses at length, there is no consensus in the literature on philosophy concerning the definition of the term. The philosophers in this volume, however, appear to share the general view that concepts and categories are grounded in experience. For Lalumera, the relations between concepts mirror the relationships between the real-world properties that they refer to. For Panaccio, concepts represent singular things. For Machery and Faucher, social control factors impact on the conditions surrounding the formation of categories and concepts. These authors consider that social categories require a more integrated approach to categorization. This may help explain why people classify humans on the basis of their physical properties; this is situated cognition with an added cultural construction.

The definitions offered by certain authors as to what constitutes a category sometimes reflect their object of study. For example, categories are defined by some as different classes of environmental situations that emerge from the interaction between the agent and the environment. This is especially the case with authors who adopt a situated or embodied view of cognition, such as Cangelosi, Harnad, Nolfi, and Poirier et al. It is also the case with others, such as Gärdenfors, for whom concepts constitute bridges between perception, reasoning, and action. In this general view, concepts and categories can be considered as abstracted from experience with the world, and they are ever-evolving rather than fixed.

In his highly detailed account of concept and category representation, Barsalou contrasts two ways of thinking about concepts. In the first one, set within a semantic memory perspective, the properties and exemplars of a category are integrated into a general description that is relatively detached from the goals of specific agents. In the second, a concept can be viewed as an *agent-dependent instruction manual* that delivers specialized packages of inferences to guide an agent’s interactions with particular category members in specific situations. For Barsalou, our conceptual system is a collection of category representations, widely distributed in the brain, and rich with knowledge acquired during one’s life span. It is a dynamic system, as it anticipates, categorizes, and provides inferences following categorization, which constitutes expertise about the world. Thus, knowledge about a category provides a great deal of detailed information about the diverse range of its instances.

A number of contributors adopt, implicitly or explicitly, a position that is in line with Barsalou’s view of concepts and categories. A strong view is expressed by Prinz, for whom concepts have their basis in perception and represent categories by reliable causal relations to category instances. In this view, concepts are not fixed but vary from occasion to occasion. In Prinz’s approach, as with most psychologists, concepts constitute the tools for categorizing. They must be built up from features, and in contrast to word-like entities, they cannot be unstructured.

Goldstone et al. share similar views and their work complements that of Barsalou, Nolfi, and Prinz. For them, category learning not only depends upon perceptual and
semantic representations but leads to the generation of these representations. In short, categorization experience not only uses descriptions of objects, it also alters these descriptions. Here again, the emphasis is on action-mediated states used as the basis for constructing and elaborating categories. Concepts are complex databases and they allow us to represent, predict, and interact with the world. Many contributors hold the view that representing and doing are intimately connected.

Other authors are more interested in categorization processes than in categories per se. This is the case of neuroscientists such as Hanson and Hanson, for whom the main focus of study is not categories or concepts as such but the study of the processes involved in categorization and the localization of these operations in the brain. Categories, however, play a central role in these studies, as they are used to test the assumptions of the category-learning systems, as is the case in Ashby and Valentin’s and in Maddox and Filoteo’s contributions. In this context, categories are user-defined, and generally represent classes of stimuli that share similar attributes (e.g., segmented lines with a particular orientation and length).

In Ravizza’s work, the notion of category is closer to that of linguists and of psycholinguists, as categorization relies on the perception of a number of distinctive features (e.g., acoustic, articulatory) that reflect the unique attributes of a particular phoneme. In addition, Ravizza holds that motor speech representations and information about acoustic features must be maintained for successful categorical perception and production.

Authors who work on data mining for categories and ontologies are rather vague about what constitutes a category or a concept. Although the aim of data mining is to extract categories and concepts from large (and sometimes noisy) databases, there is no proper definition of what exactly constitutes a category. Categories and concepts are constructs borrowed from cognitive science to help with the extraction process. They are considered to be given, and the expert user is assumed to already possess good exemplars of what to look for. In data mining, the main concern is not with categories and concepts per se, but with the operations and procedures that contribute to extracting them from large databases. Meunier et al., however, consider that it is in the selection of the labels, expressing some aspect of the “semantic” content of a class of textual entities, that the cognitive and structural dimensions of categorization are called upon in text mining.

It therefore appears that the notion of category is not uniform across disciplines and is not always a central concern. For some researchers, it is the object of study. For others, categories are the end result, and it is the process and mechanisms of categorization that are important.

3.2. The nature of categories: Discrete, vague, or other

Another topic that emerges from some of the contributions in this book pertains to the nature of categories, whether discrete, vague, or other.

Grammatical categories may be discrete or nondiscrete. Examples of discrete grammatical categories are nouns and verbs, which are defined by opposite features [see (1)]. Grammatical categories can also be mixed, in the sense that they draw some of their
properties from one category and some from another. This is the case of adjectives, which draw some of their properties from nouns and some from verbs. The features that define them in (1), [+N, +V], are a means of representing these mixed properties. Gerunds, discussed by Travis, also have mixed properties. They have the characteristics of both nouns and verbs; furthermore, there are several types of them, which differ as to how nominal or verbal they are. Grammatical categories can also be multifunctional. Some lexical items can be used with more than one function (e.g., the English word phone, which can be used as a noun – a phone – or as a verb – to phone) and have different argument structures depending on the syntactic realization (as a noun, phone has no argument structure, but as a verb, it has two arguments: the phon-er and the phon-ee) (see Travis). Multifunctional categories are also extensively documented by Gil on the basis of Riau Indonesian, where the bulk of the lexicon is claimed to be multifunctional. Multifunctional categories are also found in sign languages. Bouchard et al. report that in Quebec Sign Language, verb roots are not distinguished from noun roots and that pronouns are not distinguished from determiners. Furthermore, a great deal of Muysken’s chapter is dedicated to showing that grammatical categories are not unitary notions, but emerge at the interface of different components of human cognitive and communicative capacities.

Contributors to this book from other disciplines are concerned with the discrete nature of category representation. For Harnad, the innate ability to build discrete and hierarchically ordered representations of the environment (i.e., categories) is the basis of all higher-order cognitive abilities, including language. Categorical perception is a representational process, resulting in the compression of within-category differences between members of the same category, and the expansion of between-category distances among members of different categories, as with the categorical perception of some speech features. Categorical perception has been shown to occur in animals and human subjects, as well as in artificial systems, as shown by Nolfi and Poirier et al. McCabe et al. present novel evidence that we apprehend the world via discrete processing cycles. They suggest that the discrete partitioning of our experience with the world extends to the visual domain. This contrasts with the general view that our everyday experience of time is continuous in nature.

Categories and concepts can also be vague. Vagueness is expressed as an overlap between categories. That is, members for which it cannot be cognitively decided whether they belong in one category or another end up placed at the periphery of categories, such that adjacent categories gradually merge. But, as is argued by Diday, vagueness should not be interpreted as a weakness, and it should be distinguished from nonspecificity and ambiguity. Furthermore, graduality, or the graded membership of exemplars in a particular category, can be a useful form of vagueness as it helps to better capture the idea of typicality, and to interface linguistic categories with a continuum of attribute values without introducing arbitrary discontinuities.

3.3. Are there modality effects on categories?

Are there modality effects on categories? This question is pertinent for most contributors to the book. For linguists, modality refers to the oral or signed nature of languages.
Oral languages are primarily sequential and use the vocal/auditory channel, whereas sign languages are primarily spatial and use the manual (corporeal)/visual channel. Both Lillo-Martin and Bouchard et al. address the question of whether this difference in modality of transmission has an effect on categories. Bouchard et al. conclude that “Oral and sign languages are actually very similar in the fundamental principles of their syntax, but important physico-perceptual differences between their modalities determine the surface realizations of these principles in ways that make them appear very different.” As for Lillo-Martin, she concludes that morphosyntactic categories that appear to show the effects of the linguistic modality at first glance can be argued to be similar at a deeper level upon a second examination. Baker further explores this issue and presents a strong argument in favor of categorial similarity across modalities.

In psychology and neuroscience, and for some authors in philosophy, modality refers to the level of sensory experience (e.g., auditory, visual, tactile, etc.). Some authors see a clear association between situated and embodied cognition and a modal representation of categories. Barsalou speaks of a modal reenactment. Enactment represents the notion that, when people act, they bring structures and events into existence and set them in action. In Barsalou’s view, modal simulations underlie conceptual processing. The modal reenactment of perceptual, motor, and introspective states is assumed to be quite similar to the reenactment process underlying mental imagery. In his approach, the conceptual system shares fundamental mechanisms with modality-specific systems. Thus, in situated conceptualization, many different specialized representations can be constructed for a given concept, each tailored to situation-specific goals and constraints. This conceptual system is distributed in the brain, where “convergence zones” capture the patterns of activation evoked in the scaffolding of these representations. Barsalou’s model is an ecological conceptual system, in touch with an agent’s experience and anticipated actions. It is a dynamic system that feeds on the interactions between an agent and the environment. Although not clearly stated, the notion of emergence – well developed in Nolfi’s chapter – is an important element of this conceptual system. This general view of modal representation appears to be shared by many authors in the book, though not always explicitly.

Modality thus appears to be of central importance to most authors interested in human and artificial cognition. For linguists, modality effects on categories reduce to the visual and auditory domains. For most others, the sensorium plays an important part in the shaping of our categories and concepts.

3.4. Are there universal categories? Are there innate categories?

Several chapters of this book address the question of whether there are categories that are universal, and of whether there are innate categories. As will be seen below, people tend to hold strong views about these issues.

Like Chomsky, linguists assume that there is a language acquisition device called Universal Grammar (UG), which is part of the human biological endowment. By hypothesis, UG comprises the principles and parameters that define the form of natural languages (as opposed to artificial languages). Principles and parameters need not be
learned for they are hypothesized to be innate. What needs to be learned are the properties that are language specific: the language-specific lexical properties and the parametric values of particular languages. The problem that linguists face is distinguishing the properties of language that are universal from those that are language-specific, and those that are innate from those that are learned. For example, speaking of semantic categories, Clark questions the extent to which children’s semantic categories are informed by universal conceptual categories or by the conventions of the language community. Some conceptual categories are assumed to be part of the universal human apparatus including notions of space, event cognition, quantity/number, causality, agency, and animacy (see Papafragou). Are (some) grammatical categories universal? Competing views on this issue are presented in several chapters of this book, for example, Travis, Gil, Bouchard et al., Lillo-Martin, Baker, Labelle, and Poulin-Dubois.

Cognitive anthropologists are also concerned with distinguishing between universals and culture-specific categories, and between innate and learned ones. In the case of color perception, categorical perception is claimed to be innate rather than learned. As Harnad states, “color boundaries along the visible spectrum are a result of inborn feature detectors rather than of learning to sort and name colors in particular ways.” Furthermore, building on Berlin and Kay (1969), Cook et al. show that universal crosslinguistic constraints on color naming exist, and that basic color terminology systems tend to develop in a partially fixed order, thus revealing the universality of color naming across languages. In the case of emotions, Boster states his position in the following terms: “Rejecting both extreme universalism and relativism, I insist that emotions are at once absolutely universally biologically endowed and completely locally culturally constructed.”

Racialism has been claimed to be a by-product of a human kind module. In their chapter on why we think racially, Machery and Faucher critically examine the evidence concerning the innateness hypothesis of racialism. Based on a large number of both theoretical and empirical considerations, they reject the claim that racialism results from an innate and essentialist nature of racial categories.

Although many of the contributors in this book do not directly address the question of the innate or learned nature of categories, it is possible to find a common view between some of them. For those who consider that categories and concepts are perceptually based, and that the relations between concepts mirror the relationships between the real-world properties they refer to, the consensus would be that concepts are learned. Harnad, for one, thinks that categorization is a skill, and that, like all skills, it must be learned. Prinz makes a strong argument for the view that concepts are learned rather than innate. For authors in AI research, the construction of categories by artificial systems is definitely a product of the artificial agent’s interaction with the environment.

The view in situated and embodied conceptualization appears to be that the innate versus learned distinction is a moot point. For proponents of this position, it is possible that organisms are prepared to categorize their environment, but the perceptual basis of mental representations is what is well established.

The question of whether categories are innate or acquired is thus a concern mainly for linguistics, cognitive anthropology, and philosophy. Nearly four centuries after
Descartes revived it, the debate between the rationalists and the empiricists is still an active one. As the focus of study within some disciplines has shifted from categories to processes of categorization, this has become less of an issue for psychology and neuroscience. One issue, however, that needs further clarification is the relationship between what is universal and what is biologically endowed.

One is struck by the richness and variety of the views expressed in the book. Probably because each discipline focuses on particular objects of study, our representation of categories is still diverse, we still approach the nature of categories from different perspectives, and there is a great diversity in what constitutes the central object of study. Except in rare cases, the views expressed are not so much competing as different, and thus they hold great promise for complementarity. In the next section, we now turn to the discussion of the bridges that may be built between these views.

4. Bridging the category divide

As we saw in the preceding section, most disciplines in cognitive science approach the study of categorization from well-established perspectives. Nonetheless, there has been some effort to bring some of the contributing disciplines closer. In this section, we call attention to the bridges that have already been built, and point out avenues for further cross-fertilization between disciplines.

Among the bridges between disciplines that are on firm foundations, the neuroscience of category learning and categorization reflects the solid contribution of three founding disciplines: psychology, neuroscience, and cognitive computer science. Learning systems, brain anatomy, and neural network modeling have each contributed significantly to complementary aspects of human cognition, resulting in a better understanding of the cognitive processes involved in categorization.

Another area of study which has greatly benefited from contributions by several disciplines is grounding, recognition, and reasoning. Reflections and advances within philosophy, AI, psychology, neuroscience, and linguistics concerning concept and category formation have been well integrated. The authors refer to similar concepts and share the same vocabulary. Their contributions, as well as those on the perception of categories and inference, all converge on the importance of situated and embodied aspects of cognition. This constitutes a significant breakthrough, considering that the authors from these disciplines have reached similar conclusions from different starting points. Studies in human cognition have thus taken a turn toward more complex and ecologically valid representations of our perceptions.

Other disciplines where bridges already exist are linguistics and cognitive anthropology. Both disciplines share a nativist approach that goes back to Chomsky (1957) and Goodenough (1957). While linguists are interested in the knowledge that enables speakers of a given language to comprehend and use this language as a native speaker, cognitive anthropologists are interested in the knowledge that enables members of a society to comprehend this society and behave in it as a native. While linguists are
interested in a theory of grammar that enables speakers to learn and produce language, rather than in a simple description of their linguistic productions, cognitive anthropologists are interested in a theory of how people conceptualize the world rather than in a description of their behavior. Both disciplines are interested in discovering universals – of language and of culture, respectively – and in setting these apart from specifics – of language and of culture, respectively. It appears that the advances of the last 50 years in both linguistics and cognitive anthropology have been significant in establishing the goals and the framework for each of these disciplines.

There are other disciplines that benefit from each other’s input. This is the case of linguistics and AI, with respect to data and text mining. This is also the case with studies in the acquisition of language, where it has been shown that psychophysical distinctions impact on the acquisition of linguistic categories. In addition, philosophy and neuroscience meet in Eliasmith’s contribution on neural representations.

We now turn to potential bridges that could be established between some of the contributions (and the disciplines that they represent), and other complementary areas of study. The question of situatedness in cognition emerges as a major theme in this handbook. This general approach is certainly relevant to the study of human cognition, as it attempts to take into account all critical aspects of experience. From this point of view, progress in embodied cognition may be of value to disciplines that have mainly focused on the study of categories, and less on the input of the agent and the environment. For example, current research on grammatical categories and folk categories could benefit from advances in situated cognition.

Another area where cross-fertilization of disciplines would be welcome concerns the issue of innate perceptual or acquisitional devices. On the one hand, Boster assumes that emotions are biologically endowed. On the other, over the last decade, neuroscience has made major advances in discovering the role of cerebral structures in the processing of emotion. Explicit and implicit systems of emotions in mammalian organisms have been delineated. The issue of the biological endowment of emotions could be further informed by including neuroscientific evidence. In the same vein, Cook et al. come out in favor of the universality of color naming across languages. Again, the neuroscience literature is rich in animal and human lesion studies, and observations on the function of the visual system and the processing of color information. The discussion of the universality of color naming could benefit from biological and neuroscientific considerations concerning the capabilities of and constraints on our nervous and visual systems. Cross-fertilization might also benefit discussions of the biological endowment of UG. Linguistics, psychology, and neuroscience have made separate inroads into this general issue, with each discipline bringing its tools and methods to bear on studying this important question. Attempts to resolve this issue could be further nourished by a critical examination of research results in developmental neuroscience and psychology, developmental linguistics, and AI modeling of language acquisition systems.

These complementary considerations may not be essential for our understanding of how cognitive systems apprehend the world, but they would certainly enrich our approaches and provide other levels of foundational evidence. A welcome critical
discussion is evident in Machery and Faucher’s examination of the claimed innate propensity to classify races. They contrast psychological and social approaches to the study of racial categories and, in so doing, they introduce cognitive sociology.

References

Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley, CA).
PART 1

CATEGORIZATION IN COGNITIVE SCIENCE
Chapter 1

TO COGNIZE IS TO CATEGORIZE: COGNITION IS CATEGORIZATION

STEVAN HARNAD

Canada Research Chair in Cognitive Sciences,
Cognitive Neuroscience Center, Université du Québec à Montréal

Contents

Abstract 20
1. Sensorimotor systems 21
2. Invariant sensorimotor features (“affordances”) 21
3. Categorization 21
4. Learning 22
5. Innate categories 23
6. Learned categories 23
7. Supervised learning 24
8. Instrumental (operant) learning 24
9. Color categories 25
10. Categorical perception 25
11. Learning algorithms 26
12. Unsupervised learning 27
13. Supervised learning 27
14. Vanishing intersections? 28
15. Direct sensorimotor invariants 29
16. Abstraction and hearsay 30
17. Abstraction and amnesia 30
18. Invariance and recurrence 31
19. Feature selection and weighting 32
20. Discrimination versus categorization 32
21. Recoding and feature selection 33
22. Learned categorical perception and the Whorf hypothesis 34
23. Uncertainty reduction 35
24. Explicit learning 36
25. Categorization is abstraction 36
26. Sensorimotor grounding: direct and derivative 36
27. The adaptive advantage of language: hearsay 37

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved
Abstract

We organisms are sensorimotor systems. Things in the world come in contact with our sensory surfaces, and we interact with them based on what that sensorimotor contact “affords.” All of our categories consist of ways we behave differently toward different kinds of things – things we do or do not eat, mate with, or flee from; or the things that we describe, through our language, as prime numbers, affordances, absolute discriminables, or truths. That is all that cognition is for, and about.

Pensar es olvidar diferencias, es generalizar, abstraer.
En el abarrotado mundo de Funes no había sino detalles, casi inmediatos. Borges

("Funes el memorioso")
1. Sensorimotor systems

Organisms are sensorimotor systems. The things in the world come in contact with our sensory surfaces and we interact with them based on what that sensorimotor contact “affords” (Gibson 1979).

2. Invariant sensorimotor features (“affordances”)

To say this, is not to declare oneself a “Gibsonian” (whatever that means). It is merely to point out that what a sensorimotor system can do is determined by what can be extracted from its motor interactions with its sensory input. If you lack sonar sensors, then your sensorimotor system cannot do what a bat’s can do, at least not without the help of instruments. Light stimulation affords color vision for those of us with the right sensory apparatus, but not for those of us who are color-blind. The geometric fact that, when we move, the “shadows” cast on our retina by nearby objects move faster than the shadows of further objects means that, for those of us with normal vision, our visual input affords depth perception.

From more complicated facts of projective and solid geometry, it follows that a three-dimensional shape, such as, say, a boomerang, can be recognized as remaining the same shape – and the same size – even though the size and shape of its shadow on our retinas changes as we move in relation to it, or as it moves in relation to us. Its shape is said to be invariant under these sensorimotor transformations, and our visual systems can detect and extract that invariance, and translate it into a visual constancy. So we keep seeing a boomerang of the same shape and size even though the shape and size of its retinal shadows keep changing.

3. Categorization

So far, the affordances I have mentioned have depended on having either the right sensors – as in the case of sonar and color – or the right invariance-detectors – as in the case of depth perception and shape/size constancy. Having the ability to detect the stimulation, or to detect the invariants in the stimulation, is not trivial; this is confirmed by the fact that sensorimotor robotics and sensorimotor physiology have so far managed to duplicate and explain only a small portion of this subset of our sensorimotor capacity. But we are already squarely in the territory of categorization here, for, to put it most simply and generally, categorization is any systematic differential interaction between an autonomous, adaptive sensorimotor system and its world. It is systematic, because we do not want arbitrary interactions like the effects of the wind blowing on the sand in the desert to be counted as categorization (though perhaps there are still some inherent similarities there worth noting). Neither the wind nor the sand is an autonomous sensorimotor system; they are, jointly, simply dynamical systems, systems that interact and change according to the laws of physics.
Everything in nature is a dynamical system, of course, but some things are not only dynamical systems, and categorization refers to a special kind of dynamical system. Sand also interacts “differentially” with wind. Blow it this way and it goes this way; blow it that way and it goes that way. But that is neither the right kind of systematicity nor the right kind of differentiality. It also is not the right kind of adaptivity (though again, categorization theory probably has a lot to learn from ordinary dynamical interactions too, even though they do not count as categorization).

Dynamical systems are systems that change in time. So it is already clear that categorization too will have to have something to do with changes across time. But adaptive changes in autonomous systems are those in which internal states within the autonomous system change systematically with time, so that, to put it simply, the exact same input will not produce the exact same output across time, every time, the way it does in the interaction between wind and sand (whenever the wind blows in exactly the same direction and the sand is in exactly the same configuration). Categorization is accordingly not about exactly the same output occurring whenever there is exactly the same input. Categories are kinds, and categorization occurs when the same output occurs with the same kind of input, rather than the exact same input. And a different output occurs with a different kind of input. So that is where the “differential” comes from.

4. Learning

The adaptiveness comes in with the real-time history. Autonomous, adaptive sensorimotor systems categorize when they respond differentially to different kinds of input, but the way to show that they are indeed adaptive systems – rather than just akin to very peculiar and complex configurations of sand that merely respond (and have always responded) differentially to different kinds of input in the way ordinary sand responds (and has always responded) to wind from different directions – is to show that at one time it was not so, and that it did not always respond differentially as it does now. In other words (although it is easy to see it as exactly the opposite), categorization is intimately tied to learning.

Why might we have seen it as the opposite? Because if instead of being designers and explainers of sensorimotor systems and their capacities, we had simply been concerned with what kinds of things there are in the world, we might have mistaken the categorization problem as merely being the problem of identifying what it is that exists (that sensorimotor systems can then go on to categorize). But that is the ontic side of categories, concerned with what does and does not exist, and that is probably best left to the respective specialists in the various kinds of things there are (specialists in animals, vegetables, or minerals, to put it simply). The kinds of things there are in the world are, if you like, the sum total of the world’s potential affordances to sensorimotor systems like ourselves. But the categorization problem is not determining what kinds of things there are, but how it is that sensorimotor systems like ourselves manage to detect those kinds that they can and do detect, how they manage to respond differentially to them.
5. Innate categories

Now it might have turned out that we were all born with the capacity to respond differentially to all the kinds of things that we do respond to differentially, without ever having to learn to do so, and there are some, like Jerry Fodor (1975, 1981, 1998), who sometimes write as if they believe this is actually the case. Learning might all be trivial; perhaps all the invariances we can detect, we could already detect innately, without the need of any internal changes that depend on time or any more complicated differential interaction of the sort we call “learning.”

This kind of extreme nativism about categories is usually not far away from something even more extreme than nativism, which is the view that our categories were not even “learned” through evolutionary adaptation. That is, the capacity to categorize comes somehow prestructured in our brains in the same way that the structure of the carbon atom came prestructured from the Big Bang, without needing anything like “learning” to shape it. (Fodor’s views might well be dubbed a “Big Bang” theory of the origin of our categorization capacity.)

Chomsky has made a similar conjecture (e.g., 1976) about a very special subset of our categorization capacity, namely, the capacity to generate and detect all and only those strings of words that are grammatical according to the Universal Grammar (UG) that underlies all possible natural languages. UG compliance is the underlying invariant in question, and, according to Chomsky, our capacity to detect and generate UG-compliant strings of words is shaped neither by learning nor by evolution; it is instead somehow inherent in the structure of our brains as a matter of structural inevitability, directly from the Big Bang. This very specific theory, about UG in particular, is not to be confused with Fodor’s far more general theory that all categories are unlearned and unevolved. In the case of UG, there is considerable “poverty-of-the-stimulus” evidence to suggest that UG is not learnable by children on the basis of the data they hear and produce within the time they take to learn their first language. In the case of most of the rest of our categories, however, there is no such evidence.

6. Learned categories

All evidence suggests that most of our categories are learned. To get a sense of this, open a dictionary at random and pick out a half dozen “content” words (skipping function words such as if, not, or the). What you will find is nouns, verbs, adjectives, and adverbs all designating categories (kinds of objects, events, states, features, or actions). The question to ask yourself is: “Was I born knowing what words are and are not in these categories, or did I have to learn this?”

You can also ask the same question about proper names, even though they do not appear in dictionaries. Proper names indicate individuals (e.g., people, places) rather than kinds, but for a sensorimotor system, an individual is effectively just as much of a kind as the thing a content word designates. Whether it is Jerry Fodor or a boomerang,
my visual system still has to be able to sort out which of its shadows are shadows of Jerry Fodor and which are shadows of a boomerang. How does it do this?

7. Supervised learning

Nor is it all as easy as in that case. Consider the more famous and challenging problem of sorting newborn chicks into males and females. I am not sure whether Fodor thinks this capacity could be innate, but the “grandmaster, 8th-degree black-belt” chicken-sexers on this planet – of which there are few, most of them in Japan – say that it takes years and years of trial and error training under the supervision of masters to reach black-belt level. There are no short-cuts, and most aspirants never get past brown-belt level. (We will return to this.) Categorization, it seems, is a sensorimotor skill, though most of the weight is on the sensory part (and the output is usually categorical, i.e., discrete, rather than continuous). Like all skills, it must be learned.

So, what is learning? It is easier to say what a system does when it learns than to say how it does it. Learning occurs when a system samples inputs and generates outputs in response to them on the basis of trial and error; its performance is guided by corrective feedback. Things happen and we do something in response. If what we do is the right thing, there is one sort of consequence; if it is the wrong thing, there is another sort of consequence. If our performance shows no improvement over time, then we are like the sand in the wind. If our performance improves – more correct outputs, fewer errors – then we are learning. (Note that this presupposes that there is such a thing as an error, or miscategorization. No such thing comes up in the case of the wind blowing the sand.)

This sketch of learning should remind us of BF Skinner, behaviorism, and schedules of reward and punishment [Catania and Harnad (1988)]. For it was Skinner who pointed out that we learn on the basis of feedback from the consequences of our behavior. But what Skinner did not provide was the internal mechanism for this sensorimotor capacity that we and so many of our fellow creatures have, just as Gibson did not provide the mechanism for picking up affordances. Both these thinkers believed that providing internal mechanisms was either not necessary or not the responsibility of their discipline. They were concerned only with describing the input and the sensorimotor interactions, not how a sensorimotor system could actually do those things. So whereas they were already beginning to scratch the surface of the “what” of our categorization capacity, in input/output terms, neither thinker was interested in the “how.”

8. Instrumental (operant) learning

Let us, too, set aside the “how” question for the moment, and note that so-called operant or instrumental learning – in which, for example, a pigeon is trained to peck at one key whenever it sees a black circle and at another key whenever it sees a white circle (with food as the feedback for doing the right thing and absence of food as the feedback for
doing the wrong thing) – is already a primitive case of categorization. It is a systematic
differential response to different kinds of input, performed by an autonomous adaptive
system that responds randomly at first, but learns to adapt its responses under the guid-
ance of error-correcting feedback (thanks, presumably, to some sort of adaptive change in
its internal state).

The case of black vs. white is relatively trivial, because the animal’s sensory appa-
ratus already has those two kinds of inputs well-segregated in advance – although if,
after training on just black and white, we began to “morph” them gradually into one
another as shades of gray, and test those intermediate shades without feedback, the
pigeon would show a smooth “generalization gradient,” pecking more on the “black”
key the closer the input was to black, and more on the white key the closer the input
was to white, and approaching a level of chance performance midway between the two.
The same would be true for a human being in this situation.

9. Color categories

But if the animal had color vision, and we used blue and green as our inputs, the pattern
would be different. There would still be maximal confusion at the blue–green midpoint,
but on either side of that boundary the correct choice of key and the amount of pressing
would increase much more abruptly – one might even say “categorically” – than with
shades of gray. The reason is that between black and white there is no innate category
boundary, whereas between green and blue there is (in animals with normal blue–green
color vision). The situation is rather similar to hot and cold, where there is a neutral point
midway between the two poles, feeling neither cold nor hot, and then a relatively abrupt
qualitative difference between the “warm” range and the “cool” range in either direction.

10. Categorical perception

This relatively abrupt perceptual change at the boundary is called “categorical perception”
(CP), and in the case of color perception, the effect is innate. Light waves vary in frequency.
We are blind to frequencies above red (infrared frequencies, with a wavelength of about
800 nm) or below violet (ultraviolet frequencies, with a wavelength of about 400 nm). If
we did not have color CP, however, then the continuum from red to violet would look very
much like shades of gray, with none of those qualitative “bands” separated by neutral mix-
tures in between, which we all see in the rainbow or the spectrum.

Our color categories are detected by a complicated sensory receptor mechanism,
which is not yet fully understood, and whose components include not just light fre-
quency, but also other properties of light, such as brightness and saturation. It also
includes an internal mechanism of three specialized detectors selectively tuned to cer-
tain regions of the frequency spectrum (red, green, and blue), with a mutually inhibitory
“opponent-process” relation between their activities (red being opposed to green, and
blue being opposed to yellow). The outcome of this innate invariance extracting mechanism is that some frequency ranges are automatically “compressed.” We see them all as just varying shades of the same qualitative color. These compressed ranges are then separated from adjacent qualitative regions, also compressed, by small, boundary regions that look like indefinite mixtures, which are neutral between the two adjacent categories. And just as there is compression within each color range, there is expansion between them. Equal-sized frequency differences look much smaller and are harder to detect when they are within one color category than when they cross the boundary from one category to the other [Berlin and Kay (1969), Harnad 2003(a,b)].

Although basic color CP is inborn rather than a result of learning, it still meets our definition of categorization, because the real-time, trial-and-error process that “shaped” CP through error-corrective feedback from adaptive consequences was Darwinian evolution. Our ancestors who could make rapid, accurate distinctions based on color out-survived and out-reproduced those who could not. That natural selection served as the “error-correcting” feedback of the genetic trial-and-error variation. There are probably more lessons to be learned from the analogy between categories acquired through learning and through evolution, as well as from the specific features of the mechanism underlying color CP, but this brings us back to the “how” question raised earlier, to which we promised to return.

11. Learning algorithms

Machine-learning algorithms from artificial-intelligence research, genetic algorithms from artificial-life research and connectionist algorithms from neural-network research have all been providing candidate mechanisms for performing the “how” of categorization.

There are, in general, two kinds of learning models: so-called “supervised” and “unsupervised” ones. The unsupervised models are generally designed on the assumption that the input “affordances” are already quite salient, so that the right categorization mechanism will be able to pick them up on the basis of the shape of the input – from repeated exposure and internal analysis alone, without need of any external error-correcting feedback.

By way of an exaggerated example, if the world of shapes consisted of nothing but boomerangs and Jerry-Fodor shapes, an unsupervised learning mechanism could easily sort out their retinal shadows on the basis of their intrinsic structures alone (including their projective geometric invariants). But with the shadows of newborn chick abdomens, sorting them out as males and females would probably need the help of error-corrective feedback. Not only would the attempt to sort them on the basis of their intrinsic structural landscape alone be like looking for a needle in a haystack, but there is also the much more general problem, that the very same things can often be categorized in many different ways. It would be impossible, without error-correcting supervision, to determine which way was correct in a given context. The right categorization can vary with the context: for example, we may want to sort baby chicks sometimes by gender, sometimes by species, and sometimes by something else [Harnad (1987)].
In general, a nontrivial categorization problem will be “underdetermined.” Even if there is only one correct solution, and even if it can be found by an unsupervised mechanism, it will first require a lot of repeated exposure and processing. The figure/ground distinction might be something like this: How, in general, does our visual system manage to process the retinal shadows of real-world scenes in such a way as to sort out what is figure and what is ground? In the case of ambiguous figures such as Escher drawings, there may be more than one way to do this, but, in general, there is a default way to do it that works, and our visual systems usually manage to find it quickly and reliably for most scenes. It is unlikely that our visual systems learned to do this on the basis of having had error-corrective feedback from sensorimotor interactions with samples of endless possible combinations of scenes and their shadows.

12. Unsupervised learning

There are both morphological and geometric invariants in the sensory shadows of objects, highlighted especially when we move relative to them or vice versa. These invariants can be extracted by unsupervised learning mechanisms that sample the structure and the correlations (including covariance and invariance under dynamic sensorimotor transformations). Such mechanisms cluster things according to their structural similarities and dissimilarities, enhancing both the similarities and the contrasts. An example of an unsupervised contrast-enhancing and boundary-finding mechanism is “reciprocal inhibition,” in which activity from one point in visual space inhibits activity from surrounding points and vice versa. This internal competition tends to bring into focus the structure inherent in and afforded by the input [Hinton and Sejnowsky (1999)].

13. Supervised learning

This kind of unsupervised clustering based on enhancing structural similarities and correlations will not work, however, if different ways of clustering the very same sensory shadows are correct, depending on other circumstances (context-dependent categorization). To sort this out, supervision by error-corrective feedback is also needed; the sensorimotor structure and its affordances alone are not enough. We might say that supervised categories are even more underdetermined than unsupervised ones. Both kinds of category are underdetermined, because the sensory shadows of their members are made up of a high number of dimensions and features – their possible combinations yielding an infinity of potential shadows, making the subset of them that will afford correct categorization hard to find. But supervised categories have the further difficulty that there are many correct categorizations (sometimes an infinite number) for the very same set of shadows.
If you doubt this, open a dictionary again, pick a content word, for example, “table,” and then think of the actual thing (in this case, a table), and think of all the other things you could have called it (thing, object, vegetable, handiwork, furniture, hardwood, Biedermeyer, even “Charlie”). The other names you could have given it correspond to other ways you could have categorized it. Every category has both an “extension” (the set of things that are members of that category) and an “intension” (the features that make things members of that category rather than another category). Not only are all things members of an infinite number of different categories, but each of their features and each combination of features is a potential basis (affordance) for assigning the thing to still more categories. So far, this is, again, just ontology. But if we return to sensory inputs, and the problem facing the theorist trying to explain how sensorimotor systems can do what they do, then sensory inputs are the shadows of a potentially infinite number of different kinds of things. Categorization is the problem of sorting them correctly, depending on the demands of the situation.

Supervised learning can help. If unsupervised learning (“mere exposure”) cannot find the winning features, perhaps feedback-guided, trial-and-error training will do it, as with the pigeon’s black/white sorting and the chicken-sexing. There are some supervised learning algorithms that are so powerful that they are guaranteed to find the “needle in the haystack,” no matter how undetermined it is. This is true only as long as the feature is just underdetermined, not indeterminate (like the exact midpoint between black and white) or NP-complete – and as long there is enough data and feedback and time [as for the language-learning child, there is not, hence the “poverty of the stimulus”; Wexler (1991)]. Our categorization algorithms have to be able to do what we can do. So, if we can categorize a set of inputs correctly, then those inputs must not only have the features that can afford correct categorization, but there must also be a way to find and use those affordances. Figure 1 shows how a supervised neural net learns to sort a set of forms into three categories by compressing and separating their internal representations in hidden unit space [Tijsseling and Harnad (1997)].

14. Vanishing intersections?

Fodor and others have sometimes suggested otherwise. They have suggested that one of the reasons most categories can be neither learned nor evolved (and hence must be “innate” in some deeper sense than merely being a Darwinian adaptation) is the “vanishing intersections” problem. If you go back to the dictionary again, pick some content words, and then look for the “invariance” shared by all the sensory shadows of just about any of the things designated by those words. You will find there is none. Their “intersection” is empty. What do all the shadows of boomerangs or tables – let alone Jerry Fodors or chicken-bottoms – have in common (even allowing dynamic sensorimotor interactions with them)? And if that doesn’t convince you, then what is the sensory shadow of categories like “goodness,” “truth,” or “beauty”?

15. Direct sensorimotor invariants

There is no reason for invariance theorists to back down from this challenge. First, it has to be pointed out that since we do manage to categorize correctly all those things designated by our dictionaries, there is indeed a capacity of ours that needs to be accounted for (see Appendix A). To say that these categories are “innate” – in a Cartesian, Platonic, or cosmogonic sense, rather than just a Darwinian sense – is simply to say that they are an unexplained, unexplainable mystery. So let us reject that. Let us assume that if organisms can categorize, then there must be a sensorimotor basis for that skill, and its source must be either evolution, learning, or both. This means that there must be enough in those shadows to afford all of our categorization capacity.
16. Abstraction and hearsay

Does it all have to be a matter of direct sensorimotor invariants, always? No, but the path to goodness, truth, and beauty requires us to trace the chain of abstraction that takes us from categories acquired through direct sensory experience to those acquired through linguistic “hearsay” [Cangelosi and Harnad (2001)].

Let us consider the five sensorimotor ways we can interact differentially with things, that is, the five kinds of things we can do with things: We can see them, recognize them, manipulate them, name them or describe them. “Manipulate,” in a sense, already covers all five ways we can interact, because manipulating is something we do with things. Let us, however, reserve the word “manipulate” for our more direct physical interactions with things, animals, or people – interactions such as touching, lifting, pushing, building, destroying, eating, mating, and fleeing. “Naming” them and “describing” them is also a thing we do with them, but let us not subsume those two acts under manipulation. “Seeing” and “recognizing” are likewise things we do with things, but these too are better treated separately, rather than as forms of manipulation. And “seeing” is meant to stand in for all modes of sensory contact with things (hearing, smelling, tasting, touching), not just vision.

“Recognizing” is special, because it is not just a passive sensory event. When we recognize something, we see it as a kind of thing (or an individual) that we have seen before. And it is a small step from recognizing a thing (as a kind or an individual) to giving it a name. “Seeing” requires sensorimotor equipment, but “recognizing” requires more. It requires the capacity to abstract. To abstract is to single out some subset of the sensory input, and ignore the rest. For example, we may see many flowers in a scene, but we must abstract to recognize that some of them are primroses. Of course, seeing them as flowers is itself abstraction. Even distinguishing a figure from the ground is an abstraction. Is any sensorimotor event not abstraction?

17. Abstraction and amnesia

To answer this question, let us turn to fiction. Borges, in his 1944 short story, “Funes the Memorious,” describes a person who cannot abstract. One day, Funes fell off his horse, and from then onward he could no longer forget anything. He had an infinite rote memory. Every successive instant of his life experience was stored forever; he could mentally replay the “tapes” of his daily experience afterwards, and it would take even longer to keep reexperiencing them than it had to experience them in the first place. His memory was so good that he gave proper names or descriptions to all the numbers – “Luis Melián Lafinur, Olimar, azufre, los bastos, la ballena, el gas, la caldera, Napoléon, Agustín de Vedía” – from 1 all the way up to enormous numbers (see Appendix B). Each was a unique individual for him. But, as a consequence, he could not do arithmetic; he could not even grasp the concepts of counting and number. The same puzzlement accompanied his everyday perception. He could not understand why people with ordinary, frail memories...
insisted on calling a particular dog, at a particular moment, in a particular place, in a particular position, by the same name that they called it at another moment, in a different time, place, and position. For Funes, every instant was infinitely unique, and different instants were incomparable and incommensurable.

Funes’s infinite rote memory was hence a handicap, not an advantage. He was unable to forget. Yet, selective forgetting, or at least selective ignoring, is what is required in order to recognize and name things. Strictly speaking, a true Funes could not even exist, or if he did, he could only be a passive sensorimotor system, buffeted about by its surroundings (like the sand by the wind). Borges portrayed Funes as having difficulties in grasping abstractions, yet if he had really had the infinite memory and the incapacity for selective forgetting that Borges ascribed to him, Funes should have been unable to speak at all, for our words all pick out categories based on abstraction. He should not have been able to grasp the concept of a dog, let alone any particular dog, or anything else, whether an individual or a kind. He should have been unable to name numbers, even with proper names, for a numerosity (or a numeral shape) is itself an abstraction. There should be the same problem of recognizing either a numerosity or numeral as being the same numerosity (numeral) on another occasion as there was in recognizing a dog as the same dog, or as a dog at all.

18. Invariance and recurrence

Funes was a fiction, but Luria described a real person who had handicaps that went in the same direction, though not all the way to having an infinite rote memory. In “The mind of a mnemonist” (1968), Luria describes a stage memory-artist, “S,” whom he had noticed when S was a journalist, because S never took notes. S did not have an infinite rote memory like Funes’s, but a far more powerful and persistent rote memory than a normal person. When he performed as a memory artist, he would memorize long strings of numbers that he had heard only once, or all of the objects in the purse of an audience member. He could remember the exact details of scenes, or long sequences. He also had synaesthesia, which means that sensory events for him were richer, polysensory experiences: sounds and numbers had colors and smells; these would help him remember. But his powerful rote memory was a handicap, too. He had trouble reading novels, because when a scene was described, he would visualize a corresponding scene he had once actually seen, and soon he was lost in reliving his vivid eidetic memory, unable to follow the content of the novel. And he had trouble with abstract concepts, such as numbers, or even ordinary generalizations that we all make with no difficulty.

What the stories of Funes and S show is that living in the world requires the capacity to detect recurrences, which in turn requires the capacity to forget, or at least ignore, what makes every instant infinitely unique and, hence, incapable of exactly recurring. As noted earlier, Gibson’s (1979) concept of an “affordance” captures the requisite capacity nicely: Objects afford certain sensorimotor interactions with them. A chair affords sitting upon; flowers afford sorting by color, or by species. These affordances
are all invariant features of the sensory input, or of the sensorimotor interaction with the input, and the organism has to be capable of detecting these invariants selectively, that is, of abstracting them and ignoring the rest of the variation. If all sensorimotor features are somehow on a par, and every variation is infinitely unique, then there can be no abstraction of the invariants that allow us to recognize sameness or similarity, or identity, whether of kinds or of individuals.

19. Feature selection and weighting

Watanabe’s (1985) “Ugly Duckling Theorem” captures the same insight. He describes how, considered only logically, there is no basis for saying that the “ugly duckling” – the odd cygnet among the several ducklings in the Hans Christian Andersen fable – can be said to be any less similar to any of the ducklings than the ducklings are to one another. The only reason it looks as if the ducklings are more similar to one another than to the cygnet is that our visual system “weights” certain features more heavily than others – in other words, it is selective, it abstracts certain features as privileged. If all features are given equal weight and there are, say, two ducklings and a cygnet, in the spatial positions D1, S, D2, then although D1 and D2 do share the feature that they are both yellow, and S is not, it is equally true that D1 and S share the feature that they are both to the left of D2 spatially, a feature they do not share with D2. Watanabe pointed out that if we made a list of all the (physical and logical) features of D1, D2, and S, and we did not preferentially weight any of the features relative to the others, then S would share exactly as many features with D1 as D1 shared with D2 (and as D2 shared with S). This is an exact analogue of Borges’s and Luria’s memory effect, for the feature list is in fact infinite (it includes either/or features too, as well as negative ones, such as “not bigger than a bread-box,” not double, not triple, etc.), so unless some features are arbitrarily selected and given extra weight, everything is equally (and infinitely) similar to everything else.

But of course our sensorimotor systems do not give equal weight to all features; they do not even detect all features. And among the features they do detect, some (such as shape and color) are more salient than others (such as spatial position and number of feathers). And not only are detected features finite and differentially weighted, but our memory for them is even more finite: We can see, while they are present, far more features than we can remember afterward.

20. Discrimination versus categorization

The best illustration of this is the difference between relative and absolute discrimination that was pointed out by George Miller in his famous 1956 paper on our brains’ information-processing limits: “The magical number 7+/−2.” If you show someone an unfamiliar, random shape, and immediately afterward show either the same shape again or a slightly different shape, they will be able to tell you whether the two successive shapes
were the same or different. That is a relative discrimination, based on a simultaneous or rapid successive pairwise comparison. But if instead, you show only one of the two shapes, in isolation, and ask which of the two it is, and if the difference between them is small enough, then the viewer will be unable to say which one it is. How small does the difference have to be? The “just-noticeable-difference,” or JND, is the smallest difference that we can detect in pairwise relative comparisons. But to identify a shape in isolation is to make an absolute discrimination (i.e., a categorization), and Miller showed that the limits on absolute discrimination were far narrower than those on relative discrimination.

Let us call relative discrimination “discrimination” and absolute discrimination “categorization.” Differences have to be far greater to be able to identify what kind or individual something is than to be able to tell it apart from something else that is simultaneously present or viewed in rapid succession. Miller pointed out that if all the differences are along only one sensory dimension, such as size, then the number of JNDs we can discriminate is very large, and the size of the JND is very small and depends on the dimension in question. In contrast, if the object is in isolation, the number of regions along a sensory dimension for which we can categorize the object is approximately seven. If we try to subdivide any dimension more finely than that, categorization errors grow.

This limit on categorization capacity has its counterpart in memory too: If we are given a string of digits to remember we—unlike Luria’s S, who can remember a very large number of them—can recall only about 7. If the string is longer, errors and interference grow.

21. Recoding and feature selection

Is there any way to increase our capacity to make categorizations? One way is to add more dimensions of variation; presumably this is one of the ways in which S’s synaesthesia helped him. But even higher dimensionality has its limits, and never approaches the resolution power of the JND of sensory discrimination.

Another way of increasing memory is by recoding. Miller showed that if we have to remember a string of 0’s and 1’s, then remembering a string of seven items is about our limit. But if we first learn to recode the digits into, say, triplets in binary code, using their decimal names—so that 001 is called one, 010 is called two, and 011 is called three, etc., and we overlearn that code, so that we can read the strings automatically in the new code, then we can remember three times as many of the digits. The seven-item limit is still there, but it is now operating on the binary triplets into which we have recoded the digits: 101 is no longer three items: it is recoded into one “chunk,” called five. We have learned to see the strings in terms of bigger chunks—and it is these new chunks that are now subject to the seven-item limit, not the single binary digits.

Recoding by overlearning bigger chunks is a way to enhance rote memory for sequences, and something similar operates at the level of features of objects: Although the number of features our sensory systems can detect in an object is not infinite, it is large enough so that if we see two different objects that share one or a few features, we will not necessarily be able to detect that they share features, and, hence, that they are
the same kind of object. This is again a symptom of the “underdetermination” mentioned earlier, and is related to the so-called “credit assignment problem” in machine learning: How to find the winning feature or rule among many possibilities [Sutton (1984)]?

To be able to abstract the shared features, we need supervised categorization training (also called “reinforcement learning”), with trial and error as well as corrective feedback based on a large and representative enough sample to allow our brains to solve the credit-assignment problem and abstract the invariants underlying the variation. The result, if the learning is successful, is that the inputs are recoded, just as they are in the digit string memorization; the features are reweighted. The objects that are of the same kind, because they share invariant features, are consequently seen as more similar to one another; and objects of different kinds, not sharing the invariants, are seen as more different.

This within-category enhancement of perceived similarity and between-category enhancement of perceived differences is again the CP described earlier in the case of color. The sensory “shadows” of light frequency, intensity, and saturation were recoded and reweighted by our evolved color receptors, so as to selectively detect and enhance the spectral ranges that we consequently see as red, yellow, etc.

22. Learned categorical perception and the Whorf hypothesis

When CP is an effect of learning, it is a kind of a “Whorfian effect.” Whorf (1956) suggested that how objects look to us depends on how we sort and name them. He cited colors as an example of how language and culture shape the way things look to us, but the evidence suggests that the qualitative color boundaries along the visible spectrum are a result of inborn feature detectors, rather than of learning to sort and name colors in particular ways. Learned CP effects do occur, but they are subtler than color CP, and can only be demonstrated in the psychophysical laboratory [Goldstone (1994, 2001), Livingston, Andrews and Harnad (1998)].

Figure 2 illustrates this for a task in which subjects learned texture categorization. For an easy categorization task, there was no difference before and after learning, but for a hard task, learning caused within-category compression and between-category separation [Pevtzow and Harnad (1997)].

Yet learned CP works in much the way that inborn CP does: Some features are selectively enhanced, while others are suppressed, thereby bringing out the commonalities underlying categories or kinds. This works like a kind of input filter, siphoning out the categories on the basis of their invariant features, and ignoring or reducing the salience of noninvariant features. The supervised and unsupervised learning mechanisms discussed earlier have been proposed as the potential mechanisms for this abstracting capacity, with sensorimotor interactions also helping us to converge on the right affordances, resolving the underdetermination and solving the credit-assignment problem.

Where does this leave the concrete/abstract distinction and the vanishing-intersections problem, then? In what sense is a primrose concrete and a prime number abstract? And how is “roundness” more abstract than “round,” and “property” more abstract still?
Identifying any category is always based on abstraction, as the example of Funes shows us. To recognize a wall as a wall rather than, say, a floor, requires us to abstract some of its features, of which verticality, as opposed to horizontality, is a critical one here (and sensorimotor interactions and affordances obviously help narrow the options). But in the harder, more underdetermined cases like chicken-sexing, what determines which features are critical? (The gist of this underdetermination is there in the Maine joke: “How’s your wife?” “Compared to what?”)

23. Uncertainty reduction

Although categorization is an absolute judgment, in that it is based on identifying an object in isolation, it is relative in another sense. The invariant features that need to be selectively abstracted depend entirely on what the alternatives are, amongst which the isolated object needs to be sorted. You might ask, “Compared to what?” The answer is that the invariance is relative to the variance. Information, as we learn from formal information theory, is something that reduces the uncertainty among alternatives. So, when we learn to categorize things, we are learning to sort the alternatives that might be confused with one another. Sorting walls from floors is rather trivial, because the affordance difference is so obvious already, but sorting the sex of newborn chicks is
harder, and it is even rumored that the invariant features are ineffable in that case: They
cannot be described in words. That is why the only way to learn them is through the
months or years of trial-and-error reinforcement training guided by feedback under the
supervision of masters.

24. Explicit learning

But let us not mistake the fact that it is difficult to make them explicit verbally for the
fact that there is anything invisible or mysterious about the features underlying chicken-
sexing – or any other subtle categorization. Biederman did a computer analysis of new-
born chick abdomens and identified the winning invariants described in terms of his
“geon” features [Biederman and Shiffrar (1987)]. He was then able to teach the features
and rules through explicit instruction to a sample of novices, so that within a short time
they were able to sex chicks at the brown-belt level, if not the black-belt level. This
progress should have taken them months of supervised trial-and-error training, accord-
ing to the grandmasters.

So if we accept that all categorization, great and small, depends on selectively
abstracting some features and ignoring others, then all categories are abstract. Only
Funes lives in the world of the concrete, and that is the world of mere passive experi-
mental flow from one infinitely unique instant to the next (like the sand in the wind). To
do anything systematic or adaptive with the input would require abstraction – whether
innate or learned: the detection of the recurrence of a thing of the same kind.

25. Categorization is abstraction

What about degrees of abstractness? (Having, with G. B. Shaw, identified categoriza-
tion’s profession – abstraction – we are now merely haggling about the price.) When I
am sorting things as instances of a round-thing and a non-round-thing, I am sorting
things. This thing is round. That thing is non-round. When I am sorting things as
instances of roundness and non-roundness, I am sorting features of things. Or rather, the
things I am sorting are features (also known as properties, when we are not just speak-
ing about them in a sensorimotor sense). And features themselves are things too: round-
ness is a feature, whereas an apple is not (although any thing, even an apple, can also
be a part, hence a feature, of another thing).

26. Sensorimotor grounding: direct and derivative

In principle, all this sorting and naming could be applied directly to sensorimotor
inputs; but much of the sorting and naming of what we consider more abstract things,
such as numbers, is applied to symbols rather than to direct sensorimotor interactions
with objects. I name or describe an object, and then I categorize it. For example, I can say, “A number is an invariant numerosity” (ignoring the variation in the kinds or individuals involved). This simple proposition already illustrates the adaptive value of language: *Language allows as to acquire new categories indirectly, through “hearsay,” without having to go through the time-consuming and risky process of direct trial-and-error learning.* Someone who already knows can just *tell* me the features of an “X” that will allow me to recognize it as an “X.”

This is rather like what Biederman, along with Shiffrar (1987) did for his experimental subjects, in telling them what features to use to sex chickens, except that his method was not pure hearsay, but a hybrid method. It was show-and-tell, not just “tell,” because he did not merely *describe* the critical features verbally; he also pointed them out and illustrated them visually. He did not first pretrain his subjects on geon-naming, as Miller’s subjects were pretrained on naming binary triplets.

**27. The adaptive advantage of language: hearsay**

If Biederman had done it all with words, through pure hearsay, he would have demonstrated the full and unique category-conveying power of language. In sensorimotor learning, the abstraction usually occurs implicitly. The neural net in the learner’s brain does all the hard work, and the learner is merely the beneficiary of the outcome. The evidence for this is that people who are perfectly capable of sorting and naming things correctly usually cannot tell you *how* they do it. They may try to tell you what features and rules they are using, but as often as not their explanation is incomplete, or even just plain wrong. This is what makes cognitive science a science; for if we could all make it explicit, merely by introspecting, how it is that we are able to do all that we can do, then our introspection would have done all of cognitive science’s work (see Appendix A). In practice, we usually cannot make our implicit knowledge explicit, just as the master chicken-sexers could not. Yet what explicit knowledge we do have, we can convey to one another much more efficiently by hearsay than if we had to learn it all the hard way, through trial-and-error experience. This is what gave language the powerful adaptive advantage that it had for our species [Cangelosi and Harnad (2001), see Figure 3, Figure 3 is Plate 1.3 in the Separate Color Plate section].

Where does this leave prime numbers then, relative to primroses? It leaves them pretty much on a par, really. I, for one, do not happen to know what primroses are. I am not even sure they are roses. But I am sure I could find out, either through direct trial-and-error experience, my guesses corrected by feedback from the masters, and my internal neural nets busily and implicitly solving the credit-assignment problem for me, converging eventually on the winning invariants; or, if the grandmasters are willing and able to make the invariants explicit for me in words, I could find out what primroses are through hearsay. It cannot be hearsay all the way down, though. I will have had to learn some ground-level things the hard, sensorimotor way, if the words used by the grandmasters are to have any sense for me. The words would have to name categories that I already have.
Is it any different with prime numbers? I know they are a kind of number. I will have
to be told about factoring, and will probably have to try it out on some numbers to see
what it affords, before recognizing that some kinds of numbers do afford factoring and
others do not. The same is true for finding out what deductive proof affords, when they

Fig.3. (a). An artificial-life simulation of mushroom foragers. Mushroom-categories could be
learned in two different ways, by sensorimotor “toil” (trial-and-error learning with feedback from
the consequences of errors) or linguistic “theft” (learning from overhearing the category
described; hearsay). (b). Within a very few generations the linguistics “thieves” out-survive and
out-reproduce the sensorimotor toilers. (But note that the linguistically based categories must be
grounded in sen- sorimotor categories: it cannot be theft all the way down.) [Reproduced with
permission from Cangelosi and Harnad (2001)].
tell me more about further features of prime numbers. Numbers themselves I will have had to learn at first hand, supervised by feedback in absolutely discriminating numerosities, as provided by yellow-belt arithmeticians – for here too it cannot be hearsay all the way down. (I will also need to experience counting at first hand, and especially what “adding one” to something, over and over again, affords.)

28. Absolute discriminables and affordances

But is there any sense in which primroses or their features are “realer” than prime numbers and their features? Any more basis for doubting whether one is really more “out there” than the other? The sense in which either of them is out there is that they are both absolute discriminables: Both have sensorimotor affordances. One way I can detect these affordances is (1) implicitly, through concrete trial-and-error experience, guided by corrective feedback. (This corrective feedback need not necessarily come from a live teacher, by the way: If, for example, primroses were edible, and all other flowers toxic, or prime numerosities were fungible, and all others worthless, feedback from the consequences of the sensorimotor interactions would be supervision enough). Or, I could be guided (2) explicitly, through verbal descriptions (as long as the words used are already grounded, directly or recursively, in concrete trial-and-error experience [Harnad (1990)]). The affordances are not imposed by me; they are “external” constraints, properties of the outside world, if you like, governing its sensorimotor interactions with me. And what I do know of the outside world is only through what it affords (to my senses, and to any sensory prostheses I can use to augment them). That two plus two equals four rather than five is hence as much of a sensorimotor constraint as that projections of nearer objects move faster along my retina than those of farther ones.

29. Cognitive science is not ontology

Mere cognitive scientists (sensorimotor roboticists, really) should not presume to do ontology at all, or should at least restrict their ontic claims to their own variables and terms of art – in this case, sensorimotor systems and their inputs and outputs. By this token, whatever it is that “subtends” absolute discriminations – whatever distal objects, events, or states are the sources of the proximal projections on our sensory surfaces that afford us our capacity to see, recognize, manipulate, name, and describe them – are all on an ontological par, and subtler discriminations are unaffordable.

Where does this leave goodness, truth, and beauty, and their sensorimotor invariants? Like prime numbers, these categories are acquired largely by hearsay. The ethicists, jurists, and theologians (not to mention our parents) tell us explicitly what kinds of acts and people are good and what kind are not, and why. But the words in their explicit descriptions must themselves be grounded, either directly, or recursively, in sensorimotor invariants. Again, categories cannot be hearsay all the way down. We can also taste what
is good and what is not good directly with our senses, of course, in sampling some of their consequences. We perhaps rely more on our own sensory tastes in the case of beauty, rather than on hearsay from aestheticians or critics, though we are no doubt influenced by them and by their theories too. The categories “true” and “false” we sample amply through direct sensory experience, but there too, how we cognize them is influenced by hearsay, and, of course, the formal theory of truth looks more and more like the theory of prime numbers, with both constrained by the affordances of formal consistency.

30. Cognition is categorization

But, at bottom, all of our categories consist in the ways we behave differently toward different kinds of “things,” whether it be the “things” we do or do not eat, mate with, or flee from, or the things that we describe, through our language, as prime numbers, affordances, absolute discriminables, or truths. And, isn’t that all that cognition is for – and about?

Appendix A. There is nothing wrong with the “classical theory” of categorization

Eleanor Rosch has suggested that because we cannot state the invariant basis on which we categorize, that invariance must not exist [Rosch and Lloyd (1978)], and hence there is something wrong with the so-called “classical” theory of categorization, according to which we categorize on the basis of the invariant features that are necessary and sufficient to afford categorization.

Not only do I not think there is anything the least bit wrong with that “classical theory,” but I am pretty confident that there is no nonmagic alternative to it. Rosch’s alternative was to vacillate rather vaguely between the idea that we categorize on the basis of “prototypes,” or on the basis of “family resemblances.” Let us consider each of these candidate mechanisms in turn.

To categorize on the basis of prototypes would be to identify a bird as a bird, because it looks more like the template for a typical bird than the template for a typical fish. This would be fine if all, many, or most of the things we categorize indeed had templates, and our internal categorization mechanism could sort their sensory shadows by seeing which template they are closest to. In other words, it would be fine if such a mechanism could actually generate our categorization capacity.

Unfortunately, it cannot. Template matching is not very successful among the many candidate machine-learning models, and one of the reasons it is unsuccessful is that it is simply not the case that everything is a member of every category, to different degrees. It is not true ontologically that a bird is a fish (or a table) to a certain degree. Nor is it true functionally that sensory shadows of birds can be sorted on the basis of their degree of similarity to prototypical birds, fish, or tables. So the prototype/template theory is a non-starter as a mechanism for our categorization capacity. It might explain our typicality judgments – for example, “Is this a more typical bird than that?” – but being able to make
a typicality judgment *presupposes* being able to categorize; it does not explain it: Before I can say how typical a bird this is, I first need to identify it as a bird!

So if it is not a question of prototypes, is it about family resemblances, then? What are family resemblances? They are merely a cluster of either/or features: This is an X, if it has feature A, or B, or not C. Either/or features (disjunctive invariants) are perfectly classical (so forget about thinking of family resemblances as alternatives to classical theories of categorization). The problem is that saying that some features are either/or features leaves us no closer to answering “how” than we were before we were informed of this. Yes, some of the affordances of sensory shadows will be either/or features, but what we need to know is what mechanism will be able to find them!

The last Roschian legacy to category theory is the “basic object” level, as opposed to the superordinate or subordinate level. Here too, it is difficult to see what, if anything, we have learned from Roschian theory.

If you point to an object, say, a table, and ask me what it is, chances are that I will say, “It’s a table, rather than a Biedermeyer, or furniture, or ‘Charlie.’” So what? As mentioned earlier, there are many ways to categorize the same objects, depending on context. A context is simply a set of alternatives among which the object’s name is meant to resolve the uncertainty (in perfectly classical information-theoretic terms). So, when you point to a table and ask me what it is, I pick “table” as the uncertainty resolver in the default context. (I may imagine that the room contains one chair, one computer, one wastebasket and one table.) If I imagine that it contains four tables, I might have to identify this one as the Biedermeyer; and if there are four Biedermeyers, I may have to hope you know I’ve dubbed this one “Charlie.”

So much for subordinate levels. The same path can be taken for superordinate levels. It all devolves on the old Maine joke, which comes close to revealing a profound truth about categories: “How’s Your wife?” Reply: “Compared to what?” If we were just discussing the relative amount you should invest in furniture for your new apartment, as opposed to investing in accessories, and you forgot what was in the adjacent room and asked me what was in there (when there was just a table), I might reply, “Furniture.” If we were discussing ontology, I might say “vegetable” (as opposed to “animal” or “mineral”), etc.

So citing the “basic object level” does not help explain our categorization capacity. That is just what one arbitrarily assumes the default context of interconfusable alternatives to be, given no further information. The only sense in which “concrete” objects, directly accessible to our senses, are somehow more basic, insofar as categorization is concerned, than more “abstract” objects – such as goodness, truth or beauty – is that sensorimotor categories must be grounded in sensory experience, and the content of much of that experience is fairly similar and predictable for most members of our species.

**Appendix B. Associationism begs the question of categorization**

The problem of association is the problem of rote pairing, for example, an object with an object, a name with a name, or a name with an object. Categorization is the problem
of recognizing and sorting objects as kinds based on finding the invariants underlying sensorimotor interactions with their shadows. Associationism had suggested that this was just a matter of learning to associate tokens (instances, shadows) of an object-type with tokens of its type-name – as indeed it is, if only we can first figure out which object-tokens are tokens (shadows) of the same object-type! This is, in turn, the problem of categorization. Associationism simply bypassed the real problem, and reduced learning to the trivial process of rote association, governed by how often two tokens co-occurred (plus an unexplained influence of how “similar” they were to one another).

Some associative factors are used by contemporary unsupervised learning models, where internal co-occurrence frequencies and similarities are used to cluster inputs into kinds by following and enhancing the natural landscape of their similarities and dissimilarities. But this is internal association among representational elements and patterns (e.g., units in a neural network), not external association between input tokens. And its scope is limited, as we have seen, for most of the shadows of most of the members of most of the categories in our dictionary could not be sorted into their respective categories by unsupervised association alone, because of underdetermination. Nor is supervised learning merely rote association with the added associative cue of the category-name (as provided by the supervisory feedback). The hard work in these learning models is done by the algorithm that solves the credit-assignment problem by finding the winning invariance in the haystack – and no model can do this at human categorization-capacity scale just yet. Critics of associationism, however, drew the incorrect conclusion that because (1) we do not know what the invariance is in most cases, and (2) association is ill-equipped to find it in any case, it follows that there either is no invariance, or our brains must already know it innately in some mysterious way [Fodor (1998)].

References

Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley, CA).


Chapter 2

A MODULAR APPROACH TO GRAMMATICAL CATEGORIES
EVIDENCE FROM LANGUAGE DIVERSITY AND CONTACT

PIETER MUYSKEN

Radboud University, Nijmegen

Contents

Abstract 46
1. Introduction 47
2. Modularity and mismatch 48
3. Grammaticalization: The case of pe in Sranan (Suriname) 52
4. Mismatches in complexity of representations:
   The case of ku in Cuzco Quechua (Peru) 54
5. Lexical nondistinctness 55
   5.1. The case of timap in Palikur 55
   5.2. ‘For’ prepositions to become complementizers 56
   5.3. Adjectives versus adverbs 57
6. Lexical overspecification: Dutch gender and definiteness 57
7. The status of null elements: Subjects in Sranan and Papiamentu 58
8. Partially overlapping categories: predicate adjectives in Creoles 59
9. Differences in lexical richness 60
10. Evidence from language contact 61
    10.1. Otomanguean-Spanish language contact 62
    10.2. Borrowing of content words in Salishan languages 65
11. Concluding remarks 67
Appendix 67
References 68
Abstract

This chapter explores the light that language diversity and contact research may shed on the issue of the integrity of grammatical categories. Data are adduced from different languages, in particular Amerindian and Creole ones, to show that categories are not unitary notions, but emerge at the interface of different components of our human cognitive and communicative capacities. Special attention is paid to borrowing patterns in the Otomanguean language Popoloca. Finally, a research program is developed to reveal how typological and language contact research may illuminate these interfaces.
1. Introduction

Categorization is central to linguistics. The vast majority of linguists assume that language systems are hierarchically structured systems of categories of different types, which are manipulated as components in symbolic representation systems. The hierarchy involves something like the units listed in Table 1; units, which correspond to levels of analysis.

In this hierarchy, each unit at a particular level is assumed to consist of a combination of units from the next level down. Furthermore, on each level, the units can be uniquely classified in terms of some structured categorial system, roughly as in Table 2.

However, the sense of ordered hierarchy, of levels built up from elements that belong to a lower level, as suggested in much of structural linguistics, is very deceptive. I will argue in this chapter that the hierarchy portrayed in Tables 1 and 2 results, rather, from the interaction between different cognitive systems and does not correspond to a simple higher-than/lower-than logic. Crucial cognitive systems involved in this interaction are, at the very least:

(1) *semiotics*: yielding the units of the lexicon, words, and their shapes;
(2) *syntax*: yielding structures with category features that can be manipulated by rules.

These systems will be discussed in more detail below. Here, I will simply state that the apparently ordered hierarchy hides more complex interactions between various sub-systems. In this chapter, I will focus particularly on the relations between words as lexical categories (part of the semiotic system) and words as syntactic categories (part of the syntactic system).

Most researchers assume that the lexical categories of a language are also its syntactic categories, and that there is a one-to-one correspondence between the two. In

<table>
<thead>
<tr>
<th>Text</th>
<th>unit of a single communicative event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turn</td>
<td>unit of the part taken by a single speaker during that event</td>
</tr>
<tr>
<td>utterance</td>
<td>all directly connected uttered material within a turn</td>
</tr>
<tr>
<td>or sentence</td>
<td></td>
</tr>
<tr>
<td>clause</td>
<td>unit describing a single state of affairs or proposition</td>
</tr>
<tr>
<td>phrase</td>
<td>unit within a clause consisting of several closely linked words</td>
</tr>
<tr>
<td>word</td>
<td>minimal unit that can be uttered separately</td>
</tr>
<tr>
<td>morpheme</td>
<td>minimal unit that has meaning</td>
</tr>
<tr>
<td>phoneme</td>
<td>minimal sound component</td>
</tr>
<tr>
<td>feature</td>
<td>minimal sound specification</td>
</tr>
</tbody>
</table>
many ways, this is a reasonable assumption. Thus, if a language has adjectives in its lexicon, they can be inserted into adjective positions made available in the phrase structure rules of the language, and so on. For many linguists, it would be difficult to imagine a different setup, e.g., adjectives in the lexicon that could not be inserted, or adjective positions in the syntax that could not be filled lexically. This chapter will explore the possibility that the match between syntactic and lexical categories is not perfect, nonetheless, and in particular, will adduce evidence from language diversity and contact research, which makes the case for such mismatches even stronger.

A number of linguistic phenomena suggest the possibility of a category mismatch. These will be discussed in Sections 3–10. In Section 11, I will discuss the more general theoretical implications of the approach taken in this chapter, and suggest some avenues for further research.

### 2. Modularity and mismatch

The mismatch between syntactic and lexical categories consists in the fact that the categories characterizing the lexicon of a language need not always be relevant syntactically, and syntactically relevant categories do not always have a lexical expression. I will therefore argue that the notion of categories as single units is problematic, and that we should replace it with a multidimensional notion, holding that categorization may differ in the different mental modules that form part of the human language capacity, and that there are interface mapping functions between these different categorizations, functions which may reveal mismatches in mapping.

Let me begin with a note on the word “modular,” which may lead to some confusion. Usage #1 of the word “modular” sets grammar, particularly syntax, apart from other cognitive systems, and stresses separateness rather than interaction between language and other systems. Usage #2 stresses the fact that grammar itself may be decomposed into a number of separate modules, only some of which are strictly linguistic in character, and some of which play a role in other cognitive systems. I am following usage

<table>
<thead>
<tr>
<th>Unit</th>
<th>Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text or discourse</td>
<td>Monologue, oratory, prayer, informal conversation, etc.</td>
</tr>
<tr>
<td>Turn</td>
<td>Answer, interruption, continuation, etc.</td>
</tr>
<tr>
<td>Utterance or sentence</td>
<td>Question, command, promise, declaration, etc.</td>
</tr>
<tr>
<td>Clause</td>
<td>Finite, infinitive, small clause, etc.</td>
</tr>
<tr>
<td>Phrase</td>
<td>Noun phrase, verb phrase, etc.</td>
</tr>
<tr>
<td>Word</td>
<td>Noun, verb, adjective, adposition, adverb, pronoun, etc.</td>
</tr>
<tr>
<td>Morpheme</td>
<td>Root, prefix, suffix, bound stem, etc.</td>
</tr>
<tr>
<td>Phoneme</td>
<td>/p/, /f/, /l/, /sl/, /wl/, /yl/, etc.</td>
</tr>
<tr>
<td>Feature</td>
<td>[+ voice], [+ labial], [+ tense], etc.</td>
</tr>
</tbody>
</table>
#2: an internal modular organization but no sharp boundaries between linguistic competence and other human cognitive capacities. Figure 1 illustrates these options.

The view expounded here is related to other recent work on mapping and mismatches. Francis and Michaelis (2003, pp.1–2) describe the basis for their work on mismatches, which has resulted in a recent book, as follows: “Formal linguistic devices such as words, morphemes, and grammatical constructions are often recruited to perform semiotic functions distinct from those for which they were apparently developed, yielding forms or patterns which seem to confound otherwise discrete categories.” Mismatches are defined as follows (p. 2): “Thus we adopt the term mismatch to describe linguistic phenomena that involve a crossing of association lines in any cross-modal mapping.” Francis, who took the Autolexical Grammar framework developed by Jerry Sadock (1981) as her starting point, wants to develop a multidimensional theory of syntactic categories. She investigates the interaction of different factors – syntactic, morphological, and semantic in nature – that influence the properties of syntactic categories. An overview of her work can be found in Francis (1998), Francis and Michaelis (2003), and Yuasa and Francis (2003).

In addition to the work of Elaine Francis, it will be clear that Ray Jackendoff’s (2002) new synthesis involves a very similar set of proposals to those presented in the mismatch program of Francis and Michaelis (2003), in that he proposes a parallel architecture model. Consider the basic outline of this model, as sketched by Jackendoff (2002, p.123): “The
The overall architecture of grammar consists of a collection of generative components \( G_1, \ldots, G_n \) that create/license structures \( S_1, \ldots, S_n \), plus a set of interfaces \( I_{jk} \) that constrain the relation between structures of type \( S_j \) and structures of type \( S_k \). It is at the interfaces between the different types of structures, produced by autonomous modules or “generative components,” that the mismatches may occur. Jackendoff’s approach differs slightly from the one outlined here in that I assume that phonology is part of an autonomous semiotic component, and hence the lexicon is not only an interface module, in my view.

The most serious problem with most earlier structural approaches to language, particularly those within the structuralist and generative traditions, is that they have tended to view the language capacity as a single monolithic whole. This monolithic view corresponds to a view of linguistic categories as static objects.

Looking at the modular organization of the language capacity in more detail, we can state that in human language (at least) four essentially different systems or modules intersect. In addition to the two systems mentioned – (a) the structure-building and processing capacity (syntax), (b) the sign-forming and using capacity (semiotics) – there are (c) the capacity to engage in sustained exchanges of information (interaction), and (d) the capacity to form complex representations of information (cognition). The claim that there are four different interacting modules remains empty unless we manage to isolate the formal properties of these modules. Only the first module is specific to language. The other three play a role in many different aspects of human behavior. An overview of their properties is given in Table 3. We will examine each capacity in turn below.

A. Two crucial features of language are part of the module of syntax. These features are not found outside of language, as far as we can establish: endocentricity and movement. Endocentricity plays a role in sentence grammar (through X-bar theory), word formation (headedness), and phonology (e.g., in syllable structure). The property sometimes labeled displacement [by Chomsky (1995)] will be referred to as “movement” here (without any of the derivational claims often associated with this term): the fact that in language elements do not always appear in the place in the sequence where they are interpreted. An example is (3), in which locative \( \text{where} \) is interpreted as the location associated with \( \text{live} \):

(3) Where do you live ___? < You live in what place?

B. The semiotics module contributes a number of properties to language; these principles or properties are, however, also found in nonlinguistic semiotic systems. The first principle is distinctiveness: lexical elements must be sufficiently distinctive to contrast with other elements. A second principle is transparency: ideally, new lexical elements are transparently derived from existing elements. A third principle, elementarity, refers to the requirement that a lexical element should ideally function as a coherent whole, an atom that can be combined with other elements. This principle is often referred to as the “lexical integrity principle.” Fourth, we have the principle of analogy, which causes new forms to be built parallel to already existing forms. The principle of analogy produces lexical subsystems characterized by paradigmaticity.

C. Human interaction capacities contribute several crucial properties to language. However, these are also found outside of language. Sequentiality is a central property of
interaction, both linguistic and nonlinguistic. Through sequential patterning, information is structured and made processable. Furthermore, these sequences are marked by **cohesion** in the way elements are linked: interaction systems contain a number of cross-referencing devices to maintain the information structure throughout the sequence.

D. Finally, properties of our general **cognition** play a central role in language as well: first, there is **embeddedness**, whereby one cognitive unit is part of another one, and structures with internal hierarchy emerge. Specifically, the embedded units can be characterized by **recursion**, in that units can be simultaneously embedded in one another. Our cognitive systems function in terms of contrasts or **opposition** between different feature specifications. Finally, there is **displacement**: cognitive structures exist independently of immediate experience.

What we now consider to be the set of the unique design features of human language, including the hierarchical system of categories, is actually the result of the complex interaction between properties of entirely separate cognitive modules: syntax, semiotics, interaction, and cognition.

### Table 3
Overview of the formal properties of some of the cognitive modules involved in language

<table>
<thead>
<tr>
<th>Syntax</th>
<th>endocentricity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>“movement”</td>
</tr>
<tr>
<td></td>
<td>recursion</td>
</tr>
<tr>
<td></td>
<td>binary branching</td>
</tr>
<tr>
<td></td>
<td>?binary distinctions</td>
</tr>
<tr>
<td></td>
<td>?double articulation</td>
</tr>
<tr>
<td>Semiotics</td>
<td>vocabulary size</td>
</tr>
<tr>
<td></td>
<td>?double articulation</td>
</tr>
<tr>
<td></td>
<td>?compositionality</td>
</tr>
<tr>
<td></td>
<td>arbitrariness</td>
</tr>
<tr>
<td></td>
<td>convention</td>
</tr>
<tr>
<td></td>
<td>distinctiveness</td>
</tr>
<tr>
<td></td>
<td>transparency</td>
</tr>
<tr>
<td></td>
<td>elementarity</td>
</tr>
<tr>
<td></td>
<td>analogy</td>
</tr>
<tr>
<td></td>
<td>blocking</td>
</tr>
<tr>
<td></td>
<td>paradigmaticity</td>
</tr>
<tr>
<td></td>
<td>suppletion</td>
</tr>
<tr>
<td>Interaction</td>
<td>sequentiaility</td>
</tr>
<tr>
<td></td>
<td>cohesion</td>
</tr>
<tr>
<td></td>
<td>interactivity</td>
</tr>
<tr>
<td></td>
<td>stimulus-free, spontaneous</td>
</tr>
<tr>
<td></td>
<td>one-to-many</td>
</tr>
<tr>
<td>Cognition</td>
<td>embeddedness</td>
</tr>
<tr>
<td></td>
<td>opposition</td>
</tr>
<tr>
<td></td>
<td>displacement</td>
</tr>
</tbody>
</table>
This modular view of the language system has the crucial property that it allows us to account for the fact of language diversity. Why is there diversity at all, and what are its limits? This question may be less interesting in a purely culturalist or semiotic approach to language, in which no claim is made for a biologically conditioned human language capacity. However, when we view language from a biological perspective, the diversity we encounter is a bit of a mystery. Why do human languages not resemble each other much more than they appear to do?

Diversity may emerge because the processing systems involved in language can interact in different ways. These differences result from their respective formal properties. Differences between languages are due to differential access to features defined in other modules. I will focus here on one subtheory, that of grammaticalization: semiotic, cognitive, or interactional properties and oppositions become “visible” to syntactic operations through feature sharing at the interface. The relevant metalanguage involves notions such as visibility of features, compatibility of representations, and optimization of matching.

Grammatical categories can be defined along a number of dimensions:

(4) a. syntactic distribution: where in a clause does an item occur?
   b. semantic types and functions (arguments, predicates, etc.)
   c. types of meanings expressed (in the realm of lexical semantics)
   d. phonological properties (e.g., number of feet, stressability)
   e. derivational morphological properties: are elements morphologically complex, can they play a role in morphological processes?
   f. inflectional morphological properties (features expressed in a given form, such as plural, feminine, dative)
   g. lexical class membership (noun, verb, etc.)
   h. discourse function (linker, shifter, contrastive element, etc.)

I will argue that these dimensions are essentially independent of one another, and that apparent parallels in the classification of an item along different dimensions are essentially the result of optimization in matching. The empirical data presented in the next sections, which illustrate these different dimensions, will be drawn from language contact studies and from Amazonian languages.

3. Grammaticalization: The case of pe in Sranan (Suriname)

In the Surinam Creole language Sranan, one frequently hears questions such as:

(5) (Na) pe yu go?  
    Sranan
    FOC LOC.Q 2s go
    ‘Where are you going?’ (fieldwork data)

---

1 See the Appendix for a list of abbreviations used in the examples.
The form *pe* is ultimately derived from the English phrase *which place*, following a trajectory as in (6) [Bruyn (1995)]:

(6) Form development:
which place > uć presi > o presi > o pe > pe

Concomitantly, there has been a development in the meaning of the expression:

(7) Semantic development:
  specific question phrase
  > generalized locative question phrase
  > locative linker

The locative linker meaning shows up in locative relative clauses:

(8) Strati, *pe* son marki ibri ay santi. Sranan street LOC.Q sun mark every eye sand.
‘Streets, where the sun marks every grain of sand.’ (from a poem by Michael Slory)

Finally, we can also trace the evolution of *pe* syntactically and pragmatically:

(9) Syntactic development:
  question phrase fronted through focus
  > question word fronted through Q-word movement
  > particle

(10) Pragmatic development:
  question formation through focus
  > distinction between focused and nonfocused questions
  > blurring of the focus interpretation of questions with *na*

The particle *na* may or may not be added to the question word; originally, it had the pragmatic effect of focus, but in contemporary Sranan there is no strong focus reading with *na pe* questions.

These four parallel developments reflect the different modules of our human linguistic competence:

(11) Categories may develop along different dimensions:

<table>
<thead>
<tr>
<th>Form</th>
<th>Semiotics</th>
<th>Long &gt; Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td>Cognition</td>
<td>Concrete &gt; Abstract</td>
</tr>
<tr>
<td>Syntax</td>
<td>Computation of Features</td>
<td>Lexical (N, V) &gt; Functional (Q, T, etc.)</td>
</tr>
<tr>
<td>Pragmatics</td>
<td>Interaction</td>
<td>Intentional &gt; Automatic</td>
</tr>
</tbody>
</table>

Development along these different dimensions may (but need not) occur in parallel. The parallel development is the subject matter of grammaticalization theory, and the partial parallelism may be explained through the optimization of interface matches in the processing system.
4. Mismatches in complexity of representations: The case of \textit{ku} in Cuzco Quechua (Peru)

In the Amerindian language Cuzco Quechua, part of the large Quechua family, reflexive (12) and reciprocal (13) structures are marked with suffixes on the verb (Muysken 1979). Normally, the reflexive is added to the reciprocal, as in (13a). It is not possible to have a reciprocal by itself, as in (13b):

(12) Riku - \textit{ku} - ni. \\
see RE 1s \hfill Quechua
\textit{‘I see myself.’} (fieldwork data)

(13) a. Riku - \textit{na} - \textit{ku} - nkichis. \\
see REC RE 2p \hfill Quechua
\textit{‘You see each other.’}

b. *riku - \textit{na} - nkichis \\
see REC 2p (fieldwork data)

There is also a causative suffix, as in (14):

(14) Riku - \textit{chi} - ni. \\
see CAU 1s \hfill Quechua
\textit{‘I cause to see.’} (fieldwork data)

It is possible to combine the three affixes, as in (15a) and (15b), yielding different meanings:

(15) a. Riku - \textit{chi} - \textit{na} - \textit{ku} - nkichis. \\
see CAU REC RE 2p \hfill Quechua
\textit{‘x cause x-rec to see y’} \hfill ‘You cause each other to see someone.’

b. Riku - \textit{na} - \textit{chi} - \textit{ku} - nkichis. \\
see REC CAU RE 2p \hfill Quechua
\textit{‘x cause y to see x-rec’} \hfill ‘You each cause someone to see the other one of you.’ (fieldwork data)

It would seem that the position of \textit{-chi} is flexible with respect to that of \textit{-na} and \textit{-ku}. However, \textit{-ku} may not precede \textit{-chi}, as shown in (16a). To express the meaning of (16a), \textit{-ku} is simply absent, as in (16b):

(16) a. *riku - \textit{na} - \textit{ku} - \textit{chi} - sunkichis \\
see REC RE CAU 3-2p \hfill Quechua
\textit{‘x cause y to see y-rec’}

b. Riku - \textit{na} - \textit{chi} - sunkichis. \\
see REC CAU 3-2p \hfill Quechua
\textit{‘x cause y to see y-rec’} \hfill ‘You cause some people to see each other.’ (fieldwork data)
It turns out that -ku cannot precede -chi in simple reflexive constructions, either:

(17) a. *riku - ku - chi - nki  
    see REF CAU 2s  
    ‘x cause y to see y’

b. Riku - chi - ku - nki.  
    see CAU REF 2s  
    ‘x cause x to see y’  
    ‘You cause yourself to see x.’ (fieldwork data)

The meaning of (17a) can simply not be expressed morphologically; a periphrastic construction must be used. There is no deep morphosyntactic reason why -ku cannot precede -chi. In fact, it is possible in other varieties of Quechua (Muysken 1989). There is simply a morphophonemic restriction in Cuzco Quechua, part of the semiotic component, which states:

(18) a. Derivational suffixes containing u may only occur in final position in a string of derivational suffixes. Only one such affix may be realized with u.

b. A sequence of suffixes like ku mu pu is reduced to [kampu]

The syntactic representations involving causatives, reflexives, and reciprocals in Cuzco Quechua are quite complex, while their morphophonemic representation is quite flat and simple. Thus, there may be mismatches in the complexity of representations in different modules.

These modules are governed by principles autonomous to each module – principles defined by specific entities and concepts, such as vowel quality in morphophonemics and referential indices at the syntax-semantics interface.

5. Lexical nondistinctness

There are a number of cases, where the semiotic module does not make the same distinctions as the syntactic module: cases of lexical underdifferentiation or nondistinctness. I will discuss typical examples from Palikur, Haitian Creole, and Dutch.

5.1. The case of timap in Palikur

In the Arawakan language Palikur from French Guyana, analyzed by Launey (2004), a form like timap ‘noise, auditory perception’ may function as a verb, with either a noncausative (19) or a causative (20) meaning; it can also be an adverb (21), and a noun (22), (23).

(19) Nã-timap pak tinõ.  
    RE CIS BEN  
    ‘I listen to women singing.’
Naq-ti timap kabmanō su-kamkay.
‘The mother shouts to call her child.’

Audik weu timap.
‘A tapir walks noisily.’

Ahavuik, kadahan nawane timap.
‘In the forest, there is sometimes an echo.’

Ni amee timap kwiswa.
‘The raptor flies without noise.’ [Launey (2004)]

It is not clear how representative these examples are of the whole Palikur lexicon, but it is clear that in some languages the lexicon underdifferentiates with respect to the syntactic categories that play a role in the syntax (see Gil, this volume).

5.2. ‘For’ prepositions to become complementizers

Creole prepositions meaning ‘for’ are often related to various other categories. A typical example is the status of Haitian pou (< French pour), which can function as a preposition, complementizer, and mood marker. As a preposition, pou is used mostly as a benefactive:

Pote sa pou mwen.
‘Bring this for me.’ [Koopman and Lefebvre (1980), p. 203]

As a mood marker, pou marks obligation or irrealis:

Nou pa te pou wè sa.
‘We did not have to see this.’ [Koopman and Lefebvre (1980), p. 209]

Finally, as a complementizer, pou marks purposives or oblique complement clauses:

Li pa-jam tro ta pou chen anraje.
‘It is never too late for a dog to go mad.’ [Hall (1953), p. 192]

Confronted with this state of affairs, which is very widespread in Creole languages, creolists have adopted two approaches:

A. To assume that there are many homonyms, so that there would be three synchronically unrelated forms pou in Haitian;

B. To assume a grammar for Haitian in which the category of mood is very different from that in western European languages, in that there is no real distinction between complementizers and mood markers.

Here, a third possibility is suggested: the lexical entry is underspecified with respect to the distinctions the syntax makes. In fact, this may be a more general phenomenon
affecting prepositions. Thus, Spanish a ‘to, animate accusative’ and Afrikaans vir ‘for, to, animate accusative’ show behavior uncharacteristic of typical prepositions, in that they are deletable in certain contexts, e.g., when their complement appears in clause-initial position.

5.3. Adjectives versus adverbs

A different kind of mismatch involves cases, where languages differ in their marking of distinctions in the lexicon, though not in the syntax. Thus in English, most adjectives are marked with -ly when they occur in adverbial position, while in Dutch adjectives are generally not formally distinguished from adverbs:

(27) a. een mooi lied
   ‘a beautiful song’
   b. Zij zingt mooi.
   ‘She sings beautifully.’

Thus, the class of adverbs and the class of adjectives are not sharply delineated in the Dutch lexicon, while they often are in English (though not always, particularly in informal spoken English). Nonetheless, syntactically, the class of Dutch adjectives does not appear to behave any differently from the class of English adjectives.

6. Lexical overspecification: Dutch gender and definiteness

Sometimes the lexicon makes distinctions which play a marginal role in the syntax. Dutch nouns are characterized by a neuter/non-neuter distinction, which is expressed by the choice of the definite article (de for non-neuter versus het for neuter) and shows up in the ending of the preposed adjective in indefinite noun phrases:

(28) a. de koe ‘the cow’
   b. een witte koe ‘a white cow’
(29) a. het paard ‘the horse’
   b. een wit paard ‘a white horse’
   c. *een witte paard

The distinction is pervasive in the lexicon: you need to know the gender of any noun. However, it does not play a role elsewhere in the grammar, beyond the case of adjective agreement, as already mentioned. Andrew Carstairs-McCarthy (2000) has drawn attention in this respect to verb conjugations in many languages. It is important to know the conjugation of a verb in Romance languages to produce the right verb form, but there are no syntactic rules that refer to a conjugation.
In contrast, the definiteness distinction plays a role in Dutch syntax for noun phrases, but not in the lexicon. Dutch definite and indefinite noun phrases do not have the same distribution [see, e.g., Hoekstra (1984)]:

(30) a. Er loopt een wit paard. Dutch
   ‘There walks a white horse.’

b. *Er loopt het witte paard.
   ‘There walks the white horse.’

(31) a. Ik heb de man gisteren in Parijs gezien. Dutch
   ‘I saw the man in Paris yesterday.’

b. *Ik heb een man gisteren in Parijs gezien.
   ‘I saw a man in Paris yesterday.’

However, this distinction, which is crucial to the syntax/semantics interface, plays no role in the lexicon. In fact, definiteness is also a feature of proper nouns (32a) and personal pronouns (32b):

(32) a. Ik heb Jan gisteren in Parijs gezien. Dutch
   ‘I saw John in Paris yesterday.’

b. Ik heb haar gisteren in Parijs gezien. 21
   ‘I saw her in Paris yesterday.’

In some (many?) languages the lexicon overdifferentiates with respect to the syntactic categories that play a role in the syntax (beyond the phrase). Lexical properties are more likely to have an impact in the local domain than in the clausal domain; in other words, lexical distinctions are neutralized in higher projections.

7. The status of null elements: Subjects in Sranan and Papiamentu

Another mismatch between the lexicon and the syntax is evident in the possibility for null elements to occur in some languages as subjects or objects. The two Creole languages Sranan and Papiamentu (spoken in the Dutch Antilles, particularly on Aruba, Bonaire, and Curaçao) have many similar syntactic structures. However, with verbs like ‘seem’ Papiamentu has a null element in subject position, while Sranan has a lexical pronoun:

(33) [0] Parse ku e ta bin. Papiamentu
    [0] seem that 3s PR come
    ‘It seems that he is coming.’ (fieldwork data)

(34) A gersi dati a e kon. Sranan
    3s seem that 3s PR come
    ‘It seems that he is coming.’ (fieldwork data)
However, this difference has no known syntactic effects. Thus, extraction out of the complement subject position is blocked in both languages [see Rizzi (1982)]:

(35) *Ken parse ku ta bin? Papiamentu
  who seem that PR come
  ‘Who does it seem that is coming?’ (fieldwork data)

(36) *Suma gersi dati e kon? Sranan
  who seem that 3s PR come
  ‘Who does it seem that is coming?’ (fieldwork data)

In other words, Papiamentu behaves like a non-pro-drop language in this respect, even though it allows a null subject in some contexts [see, e.g., Muysken and Law (2001), also DeGraff (1993), who adopts a different perspective]. Significantly, however, it does not allow null subjects in argument positions:

(37) a. *ta bin Papiamentu
    PR come

     b. mi/bo/e ta bin
        1s/2s/3s PR come
        ‘I/you/(s)he comes. (fieldwork data)

8. Partially overlapping categories: predicate adjectives in Creoles

A third categorial mismatch concerns partially overlapping categories, such as adjective and predicate verb. In the Maroon Creole language Saramaccan, to mention a well-researched example, adjectives can occur both in attributive position (38a), and in predicative position (38b) (spelling is given as in the original sources):

(38) a. . . . kó gaán páu ku donú njanjá. Saramaccan
    ‘. . . becomes [a] big tree with yellow fruit.’

     b. Di boto baáka.

In predicative position, adjectives (39a) pattern with verbs (39b) and allow no copula (39c):

(39) a. di möön gaan hanse mujëë u di köndë Saramaccan
    ‘the more great beautiful (most beautiful) woman in the village’
    [Rountree (1992), p. 37]

     b. Di mujee hanse. Saramaccan
        DET woman beautiful
        ‘The woman is beautiful.’ [Alleyne (1987)]

     c. * di mujee dë hanse
Note that with reduplicated predicates the copula *de* is obligatory:

(40) Mi de tjalitjali. Saramaccan
    1s COP sad.REDUP

This suggests that the nonreduplicated forms are true verbs, and the reduplicated forms derived adjectives. Often the reduplicated form has a derived, more specific meaning than the nonreduplicated form in Saramaccan, suggesting its lexicalized status:

(41) a. Di mii bunu. Saramaccan
    DET child good
    ‘The child is good.’

b. Di mii de bunbunu.
    DET child COP good-REDUP
    ‘The child is fine.’

(42) a. A satu.
    3s salt
    ‘It has been salted.’

b. A de satusatu.
    3s COP salt.REDUP

The reduplicated forms can also occur in prenominal attributive position:

(43) di lailai goni Saramaccan
    DET load.REDUP gun
    ‘the loaded gun’

(44) di dee-dee koosu
    DET dry.REDUP cloth
    ‘the dried cloth’ [Bakker (1987), p. 25]

The lexicon/syntax mismatch arises in examples such as (38) and (39), where color terms and predicates such as *hanse* ‘beautiful’ have two overlapping distributions: as adjectives and as verbs.

9. Differences in lexical richness

Ordinarily, the major categories constitute open lexical classes. This holds for nouns in most if not all languages, and possibly for verbs. However, it does not hold fully for adpositions (whose status as a major category is subject to discussion) and, interestingly, it is not true of adjectives either. While the adjective class is open in Indo-European
languages, in languages belonging to a wide range of different language families, the class of adjectives is quite limited, and may consist of only a handful of elements [Dixon (1977)].

The cases discussed so far suggest that the categorial match between the lexicon and the syntax is not nearly as neat as is often assumed. The two systems interact, of course, but the correspondence between them is approximate rather than perfect. The mismatches between them can be classified into several types:

(a) complex phrases can be reduced phonologically, lose their concrete meanings, turn from a lexical into a functional category and lose their pragmatic force (Sranan pe in Section 4);
(b) the syntactic and semantic representations associated with a form can be much more complex than their phonological representations (Cuzco Quechua in Section 4);
(c) the same lexical form can occur in many different syntactic positions (Palikur in Section 5.1);
(d) preposition-like elements are underspecified for the syntactic position they can occupy (Haitian in Section 5.2);
(e) a similar set of syntactic categories is morphologically distinguished in one language, English, but not in another, Dutch (Section 5.3);
(f) different features are relevant in the syntax and in the lexicon (Dutch determiner phrases in Section 6);
(g) the null/non-null realization of an element has no syntactic consequences (impersonal subjects in Papiamentu and Sranan in Section 7);
(h) there are overlapping categories of adjectives and predicate verbs (Saramaccan in Section 8);
(i) what can be inserted into a major category position may or may not be an open class item (Section 9).

Thus the version of the lexicalist position that suggested that syntactic structures are simply projected from the lexicon through category-neutral X-bar rules has lost much of its appeal. It seems that syntactic structures exist separately from the lexicons, which are meant to fill these structures.

This result becomes particularly salient in the light of language contact, where the lexicon of one language is brought into contact with the syntax of another one, and lexical borrowing occurs. In the next section, I will focus on this phenomenon as yet another source of evidence for lexicon/syntax mismatches.

10. Evidence from language contact

I will discuss two very different cases of lexical borrowing here, both involving Amerindian languages: the borrowing of Spanish function words in Popoloca and the borrowing of European content words in Salishan languages.
10.1. Otomanguean-Spanish language contact

Like most of the other Amerindian languages in Mexico, languages belonging to the Otomanguean family have been in intimate contact with Spanish for several centuries at the very least. From the perspective of this chapter, the borrowing patterns in these languages are particularly interesting.

Consider the material from texts in Popoloca de Mezontla presented in Veerman-Leichsenring (1991) and recorded from Norberto Bautista Cortés (born in 1923). The Otomanguean language Popoloca is spoken by 11,200 speakers (aged 55 and older) in 11 villages near Tehuacán in the state of Puebla, Mexico; there are virtually no monolingual Mezontla speakers, and the process of language shift has been going on since the 1930s. The texts are traditional and cover the following topics: A. How pottery is made; B. How we lived before; C. Riddles (Adivinación) (more traditional); D. My grandfather’s school; E. The teacher Juan Pérez Rosales; F. How to ask for the hand of a girl. I have analyzed the Spanish elements in these texts and divided them into the broad categories of content words (Table 4) and function words (Table 5).

In Table 4, the content elements are further divided into adjectives, verbs, and nouns; the latter are subdivided into clusters according to semantic field (as are a few fixed expressions). For these semantic fields, only a few illustrative examples are given, along with the totals per text. For the verbs and adjectives, all types are given, with their distribution; for the noun categories, a few illustrative examples are provided. The number in parentheses after a category heading refers to the number of tokens.

In Table 5, functional items are listed with their distributions and their totals as well.

When one compares the two tables, several points come to the fore. The borrowed content words have a quite specific distribution, depending on the topic of the text they occur in. The number of nonconcrete words is rather limited. The total number of tokens borrowed is not much higher for the content words than for the function words (172 vs. 131), but the number of nonce-borrowings is double (45 vs. 22). Consequently, the type/token ratio for the content word borrowings is higher than for the function word borrowings: 79/172 = 0.47, as compared to 42/131 = 0.33.

Most important, however, is the bare fact that the number of function words borrowed (in tokens) is not much lower than the number of content words (tokens: 172 vs. 133; types: 79 vs. 42). In general, when we compare situations of borrowing around the world, the numbers of function words borrowed are much lower than those for content words, as is clear from the borrowability hierarchies in the recent literature [see van Hout and Muysken (1994) for an overview of the issues]. To take a simple example, Treffers-Daller (1994, pp. 99, 179) shows that in Brussels Dutch (assuming the category of function words to include adverbs except for the -ment class, juste ‘precisely,’ and adverbially used adjectives), the proportion of content words to function words is 3263/725 (four and a half times as high) on the token level, and 1008/79 (over 12 times as high) on the type level. Asymmetries between content word and function word borrowing for Strasbourg [Gardner-Chloros (1991)] and Ottawa [Poplack, Sankoff and Miller (1988)] are quite similar to, or even more extreme than, those found in Brussels Dutch.
The fact that there is this large difference between Popoloca and the other cases cited suggests that the borrowing process itself may have different features in Popoloca, at least for some categories. For interjections, the two language settings may be the same. Interjections are also a frequently borrowed category in the Brussels case: while in the

<table>
<thead>
<tr>
<th>Category</th>
<th>Total No.</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>karru</td>
<td>38</td>
<td>car</td>
</tr>
<tr>
<td>servesa</td>
<td></td>
<td>beer</td>
</tr>
<tr>
<td>platu</td>
<td></td>
<td>plate, dish</td>
</tr>
<tr>
<td>Abstract words</td>
<td>16</td>
<td>side</td>
</tr>
<tr>
<td>ladu</td>
<td></td>
<td>thing</td>
</tr>
<tr>
<td>kosa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Courtship and ritual</td>
<td>24</td>
<td>agreement</td>
</tr>
<tr>
<td>konsentimyento</td>
<td></td>
<td>Dio</td>
</tr>
<tr>
<td>Dio</td>
<td>35</td>
<td>God</td>
</tr>
<tr>
<td>Government and education</td>
<td></td>
<td>government</td>
</tr>
<tr>
<td>gubyernu</td>
<td></td>
<td>school</td>
</tr>
<tr>
<td>skwela</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural world</td>
<td>4</td>
<td>leun</td>
</tr>
<tr>
<td>klavelina</td>
<td></td>
<td>puma, lion</td>
</tr>
<tr>
<td>Social and family relations</td>
<td>23</td>
<td>señor</td>
</tr>
<tr>
<td>be:lu</td>
<td></td>
<td>sir, mister</td>
</tr>
<tr>
<td>Numbers, dates, times, measures</td>
<td>14</td>
<td>veintisiete</td>
</tr>
<tr>
<td>madrugada</td>
<td></td>
<td>27</td>
</tr>
<tr>
<td>litru</td>
<td></td>
<td>dawn</td>
</tr>
<tr>
<td>Adjectives</td>
<td>3</td>
<td>finu</td>
</tr>
<tr>
<td>potable</td>
<td></td>
<td>fine</td>
</tr>
<tr>
<td>sufisiente</td>
<td></td>
<td>drinkable</td>
</tr>
<tr>
<td>Verbs</td>
<td>15</td>
<td>tú-mantener</td>
</tr>
<tr>
<td>must</td>
<td></td>
<td>maintain</td>
</tr>
<tr>
<td>tú- súfri:_?a</td>
<td></td>
<td>suffer</td>
</tr>
<tr>
<td>tú- yuda</td>
<td></td>
<td>help</td>
</tr>
<tr>
<td>?e:- kompara</td>
<td></td>
<td>compare</td>
</tr>
<tr>
<td>tú- kompara</td>
<td></td>
<td>compose</td>
</tr>
<tr>
<td>?e:-kompone</td>
<td></td>
<td>accompany</td>
</tr>
<tr>
<td>?e:-kompaña</td>
<td></td>
<td>hand over</td>
</tr>
<tr>
<td>sentrega</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>172</td>
<td></td>
</tr>
</tbody>
</table>
Table 5
Borrowed function words in the Popoloca texts

<table>
<thead>
<tr>
<th></th>
<th>Total No.</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discourse organizers</strong></td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>pues</td>
<td></td>
<td>Then</td>
</tr>
<tr>
<td>X si fuera</td>
<td></td>
<td>...if it were</td>
</tr>
<tr>
<td>solo</td>
<td></td>
<td>Only</td>
</tr>
<tr>
<td>a ve</td>
<td></td>
<td>let’s see</td>
</tr>
<tr>
<td>sea</td>
<td></td>
<td>it were [be it?]</td>
</tr>
<tr>
<td>sí</td>
<td></td>
<td>indeed</td>
</tr>
<tr>
<td><strong>Prepositions and conjunctions</strong></td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>pero/u</td>
<td></td>
<td>but</td>
</tr>
<tr>
<td>asta</td>
<td></td>
<td>until</td>
</tr>
<tr>
<td>porke</td>
<td></td>
<td>because</td>
</tr>
<tr>
<td>ke</td>
<td></td>
<td>that</td>
</tr>
<tr>
<td>para i:ši:</td>
<td></td>
<td>for</td>
</tr>
<tr>
<td>desde</td>
<td></td>
<td>since</td>
</tr>
<tr>
<td>por</td>
<td></td>
<td>by</td>
</tr>
<tr>
<td>despwe (de)</td>
<td></td>
<td>after</td>
</tr>
<tr>
<td>lo ke</td>
<td></td>
<td>that which</td>
</tr>
<tr>
<td>sino</td>
<td></td>
<td>but rather</td>
</tr>
<tr>
<td>ya ke</td>
<td></td>
<td>Now that</td>
</tr>
<tr>
<td>de ke</td>
<td></td>
<td>of that</td>
</tr>
<tr>
<td>asta ke</td>
<td></td>
<td>Until</td>
</tr>
<tr>
<td>a</td>
<td></td>
<td>To</td>
</tr>
<tr>
<td><strong>Temporal expressions</strong></td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>ante</td>
<td></td>
<td>before</td>
</tr>
<tr>
<td>una ve(z)</td>
<td></td>
<td>Once</td>
</tr>
<tr>
<td>tyempu</td>
<td></td>
<td>Time</td>
</tr>
<tr>
<td>ora</td>
<td></td>
<td>Hour</td>
</tr>
<tr>
<td>repente</td>
<td></td>
<td>suddenly</td>
</tr>
<tr>
<td>a veces</td>
<td></td>
<td>sometimes</td>
</tr>
<tr>
<td>komo syempre</td>
<td></td>
<td>as always</td>
</tr>
<tr>
<td>syempre</td>
<td></td>
<td>Always</td>
</tr>
<tr>
<td>semana</td>
<td></td>
<td>Week</td>
</tr>
<tr>
<td>primero</td>
<td></td>
<td>at first</td>
</tr>
<tr>
<td>prinzipio</td>
<td></td>
<td>beginning</td>
</tr>
<tr>
<td><strong>Quantifiers and delimiters</strong></td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>puru</td>
<td></td>
<td>pure(ly)</td>
</tr>
<tr>
<td>tanto</td>
<td></td>
<td>so much</td>
</tr>
<tr>
<td>no mas</td>
<td></td>
<td>only</td>
</tr>
<tr>
<td>mero</td>
<td></td>
<td>merely</td>
</tr>
<tr>
<td>ni</td>
<td></td>
<td>neither</td>
</tr>
<tr>
<td>kwantu</td>
<td></td>
<td>how much</td>
</tr>
<tr>
<td>tal</td>
<td></td>
<td>such</td>
</tr>
<tr>
<td>kada</td>
<td></td>
<td>each</td>
</tr>
<tr>
<td>algo</td>
<td></td>
<td>something</td>
</tr>
<tr>
<td>mismo</td>
<td></td>
<td>same</td>
</tr>
<tr>
<td>klas</td>
<td></td>
<td>type</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>131</td>
<td></td>
</tr>
</tbody>
</table>
native lexicon interjections constitute only 5% of the tokens, in the borrowed lexicon they constitute over 12%. This may be explained by assuming that interjections are borrowed via alternational code-switching [Muysken (2000)], and that this is quite a productive route. In the Brussels case, conjunctions and prepositions are not borrowed very much (together they constitute 1% of the borrowed tokens); it is here that Popoloca differs, with 76 conjunctions and prepositions out of a total of 303 borrowed tokens (25%). I want to propose that evidence may be found here of a lexicon/syntax asymmetry.

Several arguments may be adduced for the claim that prepositions and conjunctions were also originally borrowed via alternational switching. First, of all, we sometimes have doubling:

(45) Cündä nge: theê ná ngu: karru nà para i:ši: me: ...  
    have-1p [?] that PR-look.for-1ex one car for that then ...  
    ‘We have to look for a car so that then ...’  

    thus cart-INS-1p water far for that IM-drink-1p  
    ‘Thus we carted the water from afar in order to drink it.’  

Here the Spanish preposition/conjunction para ‘for’ is combined with the Popoloca conjunction i:_i:, even though either could have been used alone [Veerman-Leichsenring (1991)]. Doubling in itself suggests a paratactic structure, typical of alternational code-mixing.

Second, notice that para is external to i:_i:, with respect to the complement clause. This external doubling is an extra argument in favor of alternation. The Spanish element is simply added or adjoined to the clause here, and adunction is always external.

However, most or all of the borrowed conjunctions and prepositions are not equivalent to Popoloca elements, as becomes clear from Veerman-Leichsenring’s grammatical description. The precise relation between constituents is often left unspecified by overt elements in Popoloca, and it seems as if the language fills empty positions in the syntax with elements from a non native lexicon. If this is correct, lexical borrowing may lead to very rapid syntactic change.

10.2. Borrowing of content words in Salishan languages

One of the problems for linguistic category research concerns the distinction between nouns and verbs (see Travis, this volume). In several North American languages, particularly in Salish, it is unclear how to distinguish nouns from verbs [Mithun (1999)]. Jelinek and Demers (1994) present data from Straits Salish illustrating the problem:

---

2 I am grateful to Michael Cysouw for earlier discussions on the topic of categorial distinctions and for useful references.
In (47), elements which in English correspond to verbs (47a), adjectives (47b), and nouns (47c) all occur in the same predicate position. This characteristic has led some researchers to conclude that those categories are not separate in Salish.

However, Van Eijk and Hess (1986) have argued that there is a clear indication in the realm of morphology of a distinction between nouns and verbs. Some affixes are limited to one of these classes; however, both occur in predicate position. This can be represented as follows:

(48) Salish

a. Nouns and verbs are distinct in the lexicon: they take different suffixes;

b. Nouns and verbs are nondistinct in the syntax: both function as predicates.

The mirror-image situation may be found in the Polynesian language Tongan, as analyzed by Broschart (1997). Here again, there is some doubt as to the categorial distinction between nouns and verbs:

(49) Tongan

a. Na’e lele e kau fefiné.
   PST run DET pl woman
   ‘The women were running.’

b. Na’e fefine kotoa e kau lelé.
   PST woman all DET pl run
   ‘The ones running were all female.’ [Broschart (1997), pp. 123–165]

What differentiates Tongan from Salish is that there is a clear difference between the verbal position (following the tense marker), and the nominal position (following the determiner and the plural marker). In addition, morphology is poor, and nouns and verbs are hard to distinguish by means of suffixes. We can summarize these facts in the following schema:

(50) Tongan

a. Nouns and verbs are nondistinct in the lexicon;

b. Nouns and verbs are distinct in the syntax: they occupy different structural positions in the clause.
In the analysis given in (48) and (50), Salish and Tongan, in occupying two opposite positions on the typological spectrum, illustrate the relative autonomy of the syntax and the lexicon. The fact that such radical mismatches are relatively rare, typologically speaking, should not daunt us. This is to be expected given that the lexicon/syntax interface is subject to optimized matching tendencies.

Where language contact research can make a contribution is by providing independent evidence of these interface mismatches. My analysis predicts (but this is no more than a prediction so far) that borrowing in Salish should be sensitive to lexical restrictions, but that any category can be borrowed in Tongan.

For Lillooet, the Salishan language studied by van Eijk (1985, p. 209), borrowing is extremely limited in any case. Only proper nouns (e.g., Bill, Mary) can be borrowed felicitously. There are not many studies on borrowing in North American Indian languages, and most focus on the cultural aspects. Huot (1948) and Bonvillain (1978) discuss Mohawk, while Prunet (1992) deals with Carrier. The available evidence suggests very strong noun–verb asymmetries, but a more comprehensive typological-comparative study is needed. The same is true of the Polynesian languages of the southern Pacific.

11. Concluding remarks

This exploratory chapter has identified a number of ways in which the different dimensions (semantic, syntactic, lexical, phonological) of a category may, but need not, match, yielding a wide array of different language types. This was illustrated particularly with Amerindian and Creole languages, and using language contact data. However, no theory of matching has yet been proposed, and such a theory is needed to make our account sufficiently precise to function as a theory of categories. Thus, interface-matching paradigms in the study of grammatical categories still have quite a long way to go. However, it is clear that a richer multidimensional notion of categories is needed to account for a range of complex facts. The simpler cases, which have given rise to the unitary notion of category in most traditional approaches, can be accounted for by a set of one-to-one mappings, and this is undoubtedly the unmarked option.

Appendix

Abbreviations used:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANT</td>
<td>anterior</td>
</tr>
<tr>
<td>1s, 2s, 3s</td>
<td>first-, second-, third-person singular</td>
</tr>
<tr>
<td>1p, 2p</td>
<td>first-, second-person plural</td>
</tr>
<tr>
<td>3-2p</td>
<td>third-person subject – second-person plural object</td>
</tr>
<tr>
<td>lex</td>
<td>first person exclusive plural</td>
</tr>
<tr>
<td>BEN</td>
<td>benefactive</td>
</tr>
</tbody>
</table>
CAU causative suffix  
CIS cislocative (toward speaker)  
COP copula  
DET determiner  
FOC focalizing particle  
IM imperative  
INS instrumental  
LOC.Q locative question particle  
MD irrealis mood  
N noun  
NEG negation  
pl plural  
PR progressive  
PST past  
Q question element  
RE reflexive suffix  
REC reciprocal suffix  
REDUP reduplication  
s singular  
T tense  
V verb

References


Gil (this volume).


Hoekstra, T. (1984), Transitivity (Foris, Dordrecht).


Rizzi, L. (1982), Italian Syntax. (Foris, Dordrecht).


Travis (this volume).


Chapter 3

PHILOSOPHICAL ANALYSIS AS COGNITIVE PSYCHOLOGY: THE CASE OF EMPTY CONCEPTS*

GEORGES REY

University of Maryland

Contents

Abstract 72
1. Introduction 73
2. Misadventures of the Classical View 73
3. Terminological issues 75
   3.1. Existential vs. (purely) intentional usage 75
   3.2. Concepts as between representations and referents 77
4. The inadequacies of Externalism 77
5. The need for internal roles 82
   5.1. The Quinean challenge 82
   5.2. The analytic data 83
   5.3. Rivals to the Analytic Explanation 85
      5.3.1. Quine 85
      5.3.2. Fodor 86
6. Conclusion 87
References 88

* This chapter is based on the opening address delivered at a conference on cognitive science held at the University of Quebec at Montreal in June 2003. A few short passages appear also in Rey (1993, 1998, and 2005). Note that square bracketed expressions refer to the concept expressed by the expression within the brackets: e.g., [GHOST] refers to the concept expressed by the term “ghost.”
Abstract

The “Classical” philosophical view of concepts as having necessary and sufficient defining conditions has fallen on hard times, supposedly both in philosophy and psychology. I will argue, however, that both its character and its motivation have been misunderstood. It is not committed to definitions being what people ordinarily use to categorize, or to their being otherwise superficially available to their minds. Nor is it committed to excluding prototypes or theoretic roles as possible definitions of concepts. What the Classical View is committed to are claims about the conceptual connections that underlie an agent’s competence to make modal judgments about what would satisfy the concept in actual and counterfactual circumstances, and related facts about what an agent finds intelligible.

Why believe there are such conceptual connections? Most discussions focus on ascriptions of concepts of existent phenomena – plants, animals, artifacts, which permit an ordinary “existential” construal of those ascriptions, and a consequent misleading reliance on actual phenomena in identifying the concept. “Empty” concepts of non-existent things – e.g., ghosts, devils – force us to a purely intentional construal, and thereby, I argue, to a consideration of concepts proper, and the need, then, of classical analyses.

On the other hand, many philosophers and psychologists have underestimated the challenges that have been raised by Quine, and most recently by Fodor, against classical analyses of concepts. I review those challenges, but conclude that, as serious as some (though not all) of them are, Quine and Fodor fail to provide adequate alternative explanations of the data that analyses would explain. Analyses seem to be needed, despite their not being as readily available as many philosophers and psychologists might have hoped. They may be as hidden as the principles of a Chomskian grammar.
1. Introduction

I was asked to present some of the present philosophical perspective on concepts. With no pretense to being able to do justice to the wide diversity of recent work, I decided to say something on behalf of the much maligned “Classical View” associated with traditional philosophy, showing how its status is really no worse, and actually may be a little better, than that of its rivals. In Section 2, I discuss the Classical View’s current misery. In Section 3, I discuss some surprisingly confusing terminological issues that are crucial to understanding any discussion of concepts, and especially to assessing the merits of the currently fashionable “Externalist” theories of conceptual content, which I go on to discuss in Section 4. Once those issues are sorted out, “empty concepts” – or concepts that apply to nothing, such as [GHOST], [PHLOGISTON] and [EUCLIDEAN LINE] – can be seen to present a serious problem for Externalism. They, as well as the standard semantic intuitions regarding conceptual connections, which I describe in Section 5, provide evidence for the internal conceptual roles sought by the Classical View that, as yet, has not been accommodated by any of its rivals.

2. Misadventures of the Classical View

Philosophical analysis has fallen on hard times. The “Classical View” of concepts on which it was based, according to which concepts have necessary and sufficient “defining” conditions, is widely thought to have been burdened by the simple failure to provide any convincing examples. Even worse, through the years, it has often been burdened by empiricists, positivists, and a surprising number of psychologists with a program of “reduction” of all concepts to sensorimotor ones. This rejection was linked to the strong “verificationist theory of meaning” that Quine (1953b) and Putnam (1962/1975) rightly pointed out was seriously at odds with good scientific practice. Wittgenstein (1953, Section 66 ff) further muddied the waters by making much of the fact that people do not seem to have available to them analyses of most of the ordinary concepts they are competent to use, but seem to know merely about ordinary usage and “family resemblances” among exemplars. This spawned several decades of research by psychologists on “prototype” and “exemplar” theories of concepts that were widely advertised as alternatives to the Classical View. And things took an even grimmer turn when Putnam (1965/1975) and Kripke (1972/1980) pointed out how the analyses of “kind” terms and concepts, like [WATER] OR [POLIO], seemed not to be of the a priori “conceptual” type philosophers might provide, but were to be supplied by scientists of the respective domains. Most recently, Jerry Fodor (1998, 2004a) has argued that belief in any kind of analytic, conceptual connections sought in the Classical View of concepts is where cognitive science – and indeed the whole of the twentieth century – “went wrong.” He joins others in advocating an “externalist” theory, whereby the

---

1 See Rey (1983/1999, 1985) for a further discussion of these issues.
content of a concept is constituted not by any role internal to a thinker, but by the
worldly phenomenon to which it bears some real causal relation (see Section 4).

Despite all this misery for the Classical View, the theme I want to pursue here is that
much of the criticism is unwarranted, arising from confusions about what the proper
role of analyses was supposed to be, and from burdening that role with conceptions of
the project of analysis that were not only needless, but independently bad ideas. I think
there are still good reasons to believe in analyses, even if philosophy may have to
engage cognitive scientists – and cognitive scientists might have to engage philosophers
– to discover them.

What is crucial to the Classical View are the conceptual connections underlying an
agent’s competence to make modal judgments with a concept: what she thinks would
satisfy a concept in various actual and counterfactual circumstances, and related facts
about what she finds intelligible. The existence of these connections is not in the least
undermined by the failure of verificationism, the superficial unavailability of analyses,
or people’s exploitation of prototypes. Whether or not someone reasons prototypically,
if on reflection she did not find the possibility of, say, a “female doctor in a black coat”
even intelligible, but as unintelligible as “married bachelor” or “round square,” that
would be a reason to think she did not really have the concept [DOCTOR] – that “doc-
tor” for her meant perhaps [TYPICAL AMERICAN DOCTOR]. This role for concepts
seems to me to be related to one of the main advantages of conceptual ascription in the
first place: it identifies what agents can understand, and how their attitudes might in
principle be modifiable by evidence and argument, abstracting semantic issues from the
epistemic ones, with which they are too often conflated. More colloquially, it factors out
the “merely verbal” issues from the genuinely substantive ones.

An Externalist like Fodor, however, might argue that questions of modality and intel-
ligibility are questions a person settles not by consulting meaning-constituting analyses,
but simply by learning a good theory of the actual phenomenon in the world to
which a concept refers. Although, for reasons I explore in Sections 3 and 4, this option
can seem plausible in the case of concepts that do genuinely refer to something real, it
is obviously problematic for empty concepts, or concepts that do not refer to any real,
or even genuinely possible phenomenon in the world – for example, [GHOST],
[DEMON], [BOGEYMAN], [PHLOGISTON], and the like. In addition, as I discuss in
Section 4, empty concepts are by no means confined to fairy tales and bad science: they
may be a crucial part of our normal, successful interaction with the world (Brentano
(1874/1973) famously argued that they were a distinctive mark of the mental). I’ll
argue, in Section 5, that a psychology of such concepts, and of the rich mental and
behavioral processes to which they give rise, requires something at least very like the
analyses that the Classical View sought.

On the other hand, there is no question that many philosophers and psychologists
have underestimated the difficulty of producing these analyses. For all my belief in
analyses, I am impressed by the skeptical challenges that have been raised against
them. Consequently, in this chapter, I want to engage in a rather delicate dialectical bal-
ancing act: on the one hand, urging the need for philosophical analysis in any
cognitive psychology, particularly one treating empty concepts; but, on the other hand, drawing attention to the formidable problems of providing such analyses. At the very least, this approach will provide some insight into how the topic presently stands in philosophy.

3. Terminological issues

Before proceeding, I would like to address some issues about surprisingly confusing terminology about concepts, which deserve far more attention than they have received.

3.1. Existential vs. (purely) intentional usage

Talk of “empty concepts” is talk of thoughts, ideas and representations that are “about nothing.” But there are two ways that a representation could be about nothing: it could be meaningless, as in the case of nonsense expressions, like “brillig.” Or it could be perfectly meaningful, but, as in the case of “Zeus,” there be no real thing in the world that it represents. These two ways of being “about nothing” give rise to a crucial ambiguity in ways of talking about “what representations represent.” On the surface “represent” would appear to express a two-place relation, as in:

(1) The word “cats” represents cats.

But this cannot be quite right, since

(2) The word “elf” represents elves

would then be false, for lack of elves: you cannot bear a real relation to something that does not exist. But there is surely a reading of (2) that makes it true, since, again, “elf” is not meaningless. It is merely “empty” (which I will confine to meaningful expressions).

So what does an “empty” term like “elf” represent? Well, it is an interesting fact that an almost universal response is that of Quine’s (1953a) fictitious philosopher, “McX”: it represents “an idea in your head.” Now, on the face of it, this would seem absurd, since (a) whatever else might be in your head, there are certainly no elves there; and, in any case, (b) if elves are actual ideas in your head, then elves would turn out to exist after all!

I will not go through all the replies that have been made to this puzzle, from Plato past Quine [see Cartwright (1960/1987)], but rather simply note that the word “represent” (and, for that matter, virtually any intentional idiom) seems to suffer from a systematic ambiguity along the following lines:

(i) If we are talking about a representation, \( x \), of some real thing \( y \), then \( x \) represents that real thing \( y \) – thus “Nixon” represents the actual man Nixon.

(ii) When there is not a real thing, as in the case of Zeus, then we rely on talk about the content of the expression \( y \) (which I will abbreviate by placing brackets around the expression, e.g. \([\text{Zeus}]\)).
I call the usage captured by (i) the “existential” usage, that captured by (ii), the “(purely) intentional” usage of “represent” (and other intentional idioms). I have expressed the second, intentional use with deliberate vagueness. It would be tempting to say “so a purely intentional use of ‘representation of y,’ for lack of any y, is really about an intentional content.” But this would not be correct, since someone thinking about Zeus and his philandering ways is not thinking about the philandering ways of an intentional content. Speaking more carefully, we should say something like: “when x represents a y that does not exist, a person is standing in the thinking relation to [y]; but this does not entail that he is thinking about [y].” Even in a purely intentional usage, “thinking about y” is one thing, whereas “thinking about [y]” quite another (see Rey (in preparation) for further discussion).

Interestingly enough, an analogous ambiguity can arise for the preposition “of” in “concept of.” For example, so long as it is believed that there is a referent to a concept, it is easy to rely on it to identify the content of the representation. Thus, we cheerfully say of the ancient Greeks, “They thought the stars were holes in the heavenly canopy,” not flinching at the extraordinary fact that this means the they did not even think stars were material objects! By contrast, I suspect we would flinch at “The ancient Greeks thought that Mercury was an angel and that Pluto ruled hell.” “Angel”? “Hell”? Did the ancient Greeks really have concepts of angels and hell? Arguably not. But their beliefs about Mercury and Hades were probably closer to our (conditional) beliefs about angels and hell (“If there were angels...”) than their beliefs about the holes in the canopy were to our beliefs about stars. Nevertheless, the ancient Greeks presumably had no concept [HELL] – and so, perhaps, they did not really have a concept [STAR] either, and we have merely relied on existential usage in saying that they did.

Some might allow that the ancient Greeks had “a very different concept of star” than we have today. This would appear to suggest, incoherently, that the ancient Greeks both did and did not share with us a concept with the content [STAR]. But arguably such talk is an attempt to combine an existential and intentional ascription to the ancient Greeks: they represented the stars, but they did this other than by using the concept [STAR]. Of course, this invites the question of why we should think that whatever concept they did have – [HOLES IN THE HEAVENLY CANOPY]? – really did represent the stars. “Concept of x,” especially used existentially, when there is a real x on which we can rely, tends to obscure this crucial question.

The distinction is, of course, close to the much-discussed distinctions between “transparent”/“de re” and “opaque”/“de dicto” readings of propositional attitudes and/or their ascriptions (see, e.g., Kaplan (1969)). I do not want to assimilate my distinction immediately to those, both because they are the objects of enough controversy on their own, and because a usual strategy –Kaplan’s– for understanding them will not work for “represent”: e.g., a transparent reading of “John thinks of Sam Clemens that he’s funny” may well involve John being related to a representation, “Mark Twain is funny,” that involves a representation, “Mark” that in fact represents Sam. But this understanding obviously can’t be available for the term “represent” itself.
3.2. Concepts as between representations and referents

Another way that this question can get obscured is by claiming that, e.g., the ancient Greeks simply deployed a different representation of the stars, much as different people might have different photographs of the stars. This accords with much psychological usage: if we are investigating a child’s “concept of a bird,” we often uncritically take for granted that the child and we are both thinking about these real things, birds, and we then concentrate on the particular way in which these birds get represented, e.g. by “prototypes” by “theories,” or by definitions, where these are understood to be items entokened in the brain. I think this, in part, is what leads psychologists, in particular, to use the word “concept” to indicate merely internal mental representations.

However, mental representations would no more be identical to concepts than are the words in a natural language. Words in a language are usually individuated syntactically, in such a way that different syntactic types, e.g. the words “city,” “metropolis,” “ville,” and “Stadt” all arguably express the same concept, as might pictures of New York, or Paris. In the case of mental processes, one person might express [number] thoughts by decimal numerals, another by binary ones; and, on some views, others might represent them by images, spatial layouts or by parts of the body. Moreover, the same representation may at different times or by different people express different concepts: an image of a body part might for one person represent a number but, for another, that very body part! Consequently, there must be something about the role of the representation, either in relation to its referent or in relation to other representations, that determines which concept it expresses. This is the notorious third “semantic” realm, over which theorists contend. I argue that philosophical analyses are required not for an account of the computational character of representations, but for an account of the conditions of conceptual competence, i.e. for someone having a representation with a specific conceptual content.

4. The inadequacies of Externalism

So what are those conditions of conceptual competence? Well, there is the “Classical View,” according to which a representation expresses a certain content that can be identified by a set of necessary and sufficient defining conditions. That is the view of “philosophical analysis” that has fallen on hard times and to which I will return after considering a recently influential alternative to it, namely, the “externalist” approach that has arisen in the last 30 years in response to the aforementioned work of Putnam and Kripke.

In a way, externalism might be said to make a principled virtue of what I have been identifying as the abovementioned reliance on merely “representations” and “referents,” dividing all the work of concepts between them. In particular, there is what I will call the Strong Externalism advocated by a great diversity of recent philosophers:

---

3 Talk of mental imagistic representation does seem to me, however, fraught with confusions. See Pylyshyn (2003) and Rey (1981, in preparation) for discussion.
e.g., Stampe, Dretske, Stalnaker, Fodor, and Millikan – and, before them, Skinner and Quine\textsuperscript{4}. To a first approximation (the details and many differences between their views will not matter here), \(x\) has the concept \(y\) if some state of \(x\) stands in the right causal relation to some actual \(y\): e.g., would causally covary with \(y\), was caused by \(y\) under certain (ideal, normal, evolutionarily significant) conditions. Thus, for example, someone has the concept [HORSE] if she could, under certain conditions, tell the horses from the non-horses; or has a state that was evolutionarily selected for responding to horses\textsuperscript{5}. The important feature of Strong Externalism is the claim that the content of a representation is constituted by the real phenomenon to which it is, one way or another, causally related. Different concepts of the same referent are simply different representations, often with different logical structure, that are causally linked to the same phenomenon.

I hasten to distinguish this Strong Externalism from Weak Externalism, according to which the content of some representations depends, in some way or another, upon some or other interesting relations the system of which they are a part bears to an external environment. This view allows that ascription of content to a system cannot, in general, be done in complete abstraction from the world in which the system happens to be embedded, but leaves wide open the specific ways in which that embedding may figure in any particular ascription. Perhaps a direct causal route between representation and referent is required for perceptual demonstratives; perhaps some sort of covariation is needed for certain predicates; and perhaps something else is required for fictional and other empty terms. I will not dispute this Weak Externalism view here. It seems to me supported by the usual examples – “twin” cases as in Putnam (1975) and Burge (1979)\textsuperscript{6}– and, indeed, is pretty much all that is supported by them: Strong Externalisms are (to my mind) tenuous theories of meaning intended to explain them. One kind of problem faced by Strong Externalism is how to specify the right causal relation without presupposing the very intentionality one is trying to explain\textsuperscript{7}. Although I suspect that it is a likely insuperable problem, I do not want to

\textsuperscript{4} See Rey (1997, ch.9) for a review of these proposals. It is worth noting that strong externalism is also taken for granted even by theorists who do not believe there are real phenomena in the world that representations represent. Thus, Chomsky (2000) emphasizes how phonetic phenomena do not really exist in the acoustic stream, but, amazingly enough, this leads him to suppose that representations do not have any intentional content at all! I suspect this extreme reaction is due in part to his presuming, with many of his opponents, that any intentionality must be strongly externalist [see Chomsky (2000, pp. 156–160) and Rey (2003) for further discussion].

\textsuperscript{5} Or, on Fodor’s (1991) ingenious “asymmetric dependency” (or “locking”) proposal, she would token ‘horse’ in the presence of non-horses only if it was a law that she would token it in the presence of horses, but not vice versa. The complexities of this account will not be an issue here.

\textsuperscript{6} These are the examples of beings that are physical duplicates, but whose thought contents seem to depend upon which environment they inhabit. The simplest case is provided by proper names: imagine a duplicate of President Bush inhabiting a planet that enjoyed precisely the same kind of physical history as the Earth; the proper name thoughts of his twin there would clearly not be representing the very same people or places in his life as Bush ordinarily does in his (the twin’s “Iraq” thoughts would refer to an entirely different place from Bush’s “Iraq” thoughts), despite their internal physical type identity.

\textsuperscript{7} See the introduction and essays in Loewer and Rey (1991), as well as Greenberg (2000) for various criticisms along these lines.
pursue it here. Rather, I want to consider the “empty concepts” that pose an obvious problem for any Strong Externalisms, even one specifying the right causal relationships. Whatever those relationships might be, they cannot be had with things that do not exist.

Fodor (1987, 1991, 1998) has been the most explicit about how to deal with such cases, making a familiar, but what seems to me excessively easy move. True, enough, there are no elves; but, he claims, there might have been – there is the (uninstantiated) property of elfhood. And, for Fodor, it is enough that there is a law that tokens of “elf” would be caused by elfhood counterfactually for the right meaning-constituting causal relation to be secured. What Fodor is exploiting here is a certain profligacy about properties that is not usually tolerated for mere material objects: for Fodor, properties can exist uninstantiated, and laws involving counterfactuals, about what would have happened if the world had been different than it is, are apparently there just for the asking.

Against this view, I would like first to enter a certain methodological protest, or anyway, a plea for honest toil: except as a last resort, do not introduce into your psychological ontology entities for which there is no independent evidence. Before we glibly resort to elfhood – or to being phlogiston, or being divine – and to laws and counterfactuals regarding them, we need to ask whether, independent of our semantic desperation, there is any reason whatsoever to believe in such things? I know of none. Our best theories of the world – biology, chemistry, physics, and cosmology – do not seriously make any room for such things or properties, or for laws or counterfactuals relating them.

Nor does anyone seriously expect them to. It is not that natural scientists might not have occasion to seriously postulate some uninstantiated properties. There may be many instances of magnitudes specified by differential equations, or isotopes, transuranic elements, or other phenomena that never happen to be instantiated in the actual history of the world, but which nonetheless need to be included as values of physical variables. It is just that ghosts, elves, and unicorns are surely not among them. Indeed, note that, revealingly, whereas we would turn to chemists and physicists to learn about the uninstantiated magnitudes or elements, we would turn not to them, but to psychologists or folklorists or historians to learn about, well, of course, not elves or ghosts themselves, but about people’s thoughts about them – reflecting precisely the switch in empty cases from existential to purely intentional usage of “represent,” which, we observed earlier (Section 3.1), directs us to thinking about concepts proper, rather than their referents.

In any case, some properties are, uncontroversially, necessarily uninstantiated, e.g., [ROUND SQUARE], [LARGEST PRIME], [RED AND NOT RED], [ROUND-SQUARE]; and if a property is necessarily uninstantiated then surely there cannot be laws with which a representation can causally interact, actually or counterfactually. Fodor (1991, 1998) briefly addresses this problem, and, observing that many examples are logically complex, articulates a famous view that goes back to Russell (1912/1959): “There can be no primitive concept without a corresponding property for it to [co-vary with].” (1998, p. 165). Fodor simply expects there is metaphysics enough for all our primitive thoughts, an instantiateable property for every primitive predicate. This seems to me to be a bit rash. Indeed, there seem to be a number of cases that would at least appear to present prima facie counterexamples.
Take, for starters, the familiar primitive concepts of Euclidean geometry that most high school students grasp, e.g., [POINT], [LINE], [PLANE], and [CIRCLE]. As many since Plato have pointed out, these concepts may well be uninstantiable. At any rate, we could never be sensorily presented with any such thing – all perceptible points and lines have some thickness, and so no representation in our head could enter into causal relations with any such thing or property.

Nor are uninstantiable primitives limited to geometrical examples. There is a long tradition of philosophers – Spinoza, Hume, Wittgenstein – who have plausibly argued that nothing could possibly satisfy such primitive “supernatural” concepts such as [SOUL], [ELF], [MIRACLE], [MAGIC], [MONSTER], [FREE WILL], [DESTINY], and [KARMA] (see Slote (1975) for discussion). And, moving away from these primitive supernatural examples, many philosophers have raised analogous worries about our “real world” concepts of [COLOR], [SOUND], [CAUSE], and [MATERIAL OBJECT].

In these latter cases, philosophers are often spontaneously joined by psychologists, who are fond of claiming that colors, sounds, phones, and phonemes don’t really exist: these things are – as they often put it, echoing McX – “merely in the mind.” In their classic discussion of English phonology, for example, Chomsky and Halle (1968) wrote:

… there is nothing to suggest that these phonetic representations also describe a physical or acoustic reality in any detail. For example, there is little reason to suppose that the perceived stress contour must represent some physical property of the utterance in a point-by-point fashion. In fact there is no evidence from experimental phonetics to suggest that these contours are actually present as physical properties of utterances in anything like the detail with which they are perceived.

—Chomsky and Halle (1968, p. 25)

And, in a recent, comprehensive text on vision, Stephen Palmer (1999) wrote:

Neither objects nor lights are actually “colored” in anything like the way we experience them. Rather, color is a psychological property of our visual experiences when we look at objects and lights, not a physical property of those objects or lights. ... Color is more accurately understood as the result of complex interactions between physical light in the environment and our visual nervous systems. ... There may be light of different wavelengths independent of an observer, but there is no color independent of an observer, because color is a psychological phenomenon that arises only within an observer.

—Palmer (1999, pp. 97–99)

If these views are correct, then even the content of the primitives of linguistic and visual perception cannot be constituted by their referents in the real world8. At any rate, examination of many of these concepts reveals that people, on reflection, are hard put to say what a world would be like in which there were real colors or Euclidean shapes,

8 Neither Chomsky and Hall (1968), nor Palmer (1999), of course, address the (peculiarly philosophical) modal question of whether phones or colors could exist in some other possible world, but the burden is surely on the externalist to show how they could exist there, especially in view of the difficulties of identifying them in the actual world. Fodor (1998) suggests a dispositional strategy, but this risks rendering his externalism circular [see Rey (in preparation) for discussion].
or in which people were relevantly “free” to do otherwise than they do, and in which souls could survive bodily death. Quite apart from the metaphysical issues, these facts are by themselves psychologically important ones that need to be explained. It is hard not to think that some of this persistent puzzlement is due to there being something inherently problematic about these concepts. If there are at least two such cases, an externalist theory will face a difficulty distinguishing between them.

Fodor deals with empty cases by being generous with the (possible) phenomena the world provides. Another externalist, Ruth Millikan (1998) takes a different, more niggardly line: she denies there are actually any genuine thoughts at all in such cases! She writes that such empty substance concepts

\[ \text{are not substance concepts at all. An ability that is not an ability to do anything is not an ability at all.} \]

Empty substance concepts result from failures of the mechanisms designed to develop substance concepts. They are “concepts” only in that their biological purpose was to have been concepts


When pressed for clarification, she insists that empty terms are not “real terms,” that they are like “an ability that is not an ability to do anything,” that it “makes no sense to talk of a mode of presentation that does not present anything,” and that, indeed, when reference “fails utterly there is no thought at all” (pc). This seems to be a generalization of a view Gareth Evans (1982) sketched regarding empty singular terms in natural language, but is much more radical in ways that I doubt he would have endorsed.

First, note that, according to my earlier terminological observations, it makes perfect sense to talk of “a mode of presentation that does not present anything”: it is just that one needs to distinguish, as I have, existential from intentional usages. More importantly, if apparent “thoughts” with empty terms are not really thoughts at all, then how on earth are we to explain the often rational, content sensitive behavior of people (and their subsystems) that seem to have these thoughts? People pray, make sacrifices, and engage in often elaborate reasonings about gods, devils, elves, angels, ghosts; the visual system seems to compute over representations of Euclidean figures and colors; and astronomers once intelligently sought evidence for the luminiferous ether and the planet Vulcan. In all such cases, empty thoughts interact in myriad inferential ways with non-empty

\[ \text{9 In reply to my (2004) challenge to him to explain why we should expect the world to so conveniently conform to our thought as to supply a property for every primitive predicate, Fodor (2004b) replied:} \]

\[ \text{What I suppose is that our ways of thinking about [the world] accord pretty well with the way the world is. That is sort of what you’d really expect. Making thoughts that accord with the world is what cognition is for.} \]

(p. 109, fn13)

What is curious about this reply is that it goes so much against the grain of his (Fodor [1987]) otherwise quite reasonable skepticism about teleological and/or selectionist accounts of psychological traits. Indeed, as examples of geometrical contents in the visual system show, a system with necessarily empty concepts may get on very well – provided, I suppose, that the errors it makes do not matter.

\[ \text{10 But even Evans’ (1982) views and the related views of, for example, Taylor (2003) and Campbell (2002) would seem to face serious \textit{prima facie} problems with our apparent ability to refer to stable, illusory “things” such as “particular” rainbows, animated cartoon characters, and Kanizsa triangles; see Rey (forthcoming a, in preparation) for further discussion.} \]
thoughts (e.g., about churches, misdeeds, light, and the sun), ways that surely require the empty thoughts to possess some kind of intentional content.

5. The need for internal roles

I do not see how we can hope to determine the psychology of such empty thoughts, particularly the necessarily empty ones, without saying something about the role of these thoughts in a person’s mind: again, if you want to know about the nature of ghosts, you look not to what people have been “getting at” in the world, or even in any genuinely possible world, but merely to what they think. [GHOST], but not [WILL], is tied to the imagined possibility of a disembodied mind; [WILL], but not [GHOST] is tied to maybe impossible claims about spontaneity and moral responsibility. Euclidean [POINT], [LINE], and [CIRCLE] form a cluster of interdefinable notions subject to certain axioms; [COLOR] and [SOUND] are subject to projections onto the world of certain stable reactions in our visual system.

Such a claim is not without some independent motivation. Internal roles are generally acknowledged to be needed for concepts of logic and mathematics such as [NOT], [ALL], and [SUCCESSOR], for personal-response dependent concepts such as [FUNNY], [CUTE], [WORTHWHILE], and perhaps even for concepts such as [BEAUTIFUL] and [GOOD], which also present problems for any purely externalist view. It is doubtful that there are real phenomena of “ands” and “nors” in the world, or real properties of being funny or being cute; and, even if there were, people with “the same concept” of, e.g., funniness might well disagree irresolvably, and so differ in their causal relations to external phenomena.

5.1. The Quinean challenge

But now for the other side of the dialectic. As pressing as the need might be to find some kind of “role” account of concepts, there are substantial difficulties in providing one. For starters, there is the fact that people seem to be able to believe most anything – that trees are ideas (Berkeley), that everything is numbers (Pythagoreans), that everything is a text (contemporary literary theorists). Moreover, as Quine (1953b) has emphasized, all beliefs seem revisable: after all, you can at least fool some of the people some of the time.

In saying that roles here are crucial, I do not mean to be denying a Weak Externalism (Section 4 above), according to which assigning any content whatsoever to a system may depend generally upon the environment in which it is embedded. Nor do I intend to deny that the content of some terms (e.g., “water”) that do manage to refer to something real acquire their content thereby. I am not here providing a theory of content, but merely denying what seem to me rash theories that have been proposed.

time about almost anything. So it would appear impossible to tie a concept to any specific role.

However, these considerations are hardly decisive. I have argued elsewhere [Rey (1993, 1998)] that Quine’s “revisability” argument, like much of the psychology of his time, suffers from an excessive “superficialism,” or the bad idea that psychologically real distinctions should be available on our behavioral or introspective surface. People can be mistaken about almost anything and revise their beliefs in the light of further experience and argument. But that does not show that there are not deeper inferential principles governing their thought that might belie these superficial errors. There is all the difference in the world between the clever Bishop Berkeley, who maintains all objects are ideas, and someone genuinely confused, who really does think without further ado that objects do not persist unperceived. Berkeley, after all, did feel obliged to enlist God for the tree in the lonely forest, exhibiting his competence with [MATERIAL OBJECT], despite his perverse performance. He is simply like someone who deliberately violates some rule of grammar (which is not to say there might not be good reasons for them – or for Berkeley – to do so).

5.2. The analytic data

The analogy with grammar seems to me quite close. As with grammar, we are confronted by regularities that need to be explained. In the kinds of cases that the analytic is supposed to explain, these include, inter alia, a non-negligible convergence in people’s judgments with regard to particular concepts and/or lexical items, e.g., about synonymies, ambiguities, antinomies, entailments. Ask speakers whether “bank,” or “pride” are ambiguous; whether “killings” necessarily involve “dyings;” whether “open” and “closed” are antonyms; whether “Jim boiled the water” implies “The water boiled;” whether “The square-root of a shadow is red” is anomalous. Ask people what things qualify as a bachelor, knowledge, coercion, a voluntary act, perception, water, and brisket; elicit their considered judgments about hypothetical cases; ask them what they think constitutes being one of these things: I submit that one finds what philosophers have found since Socrates, that there is great deal of convergence about what clearly does and what clearly does not satisfy a concept, and that, although people cannot readily provide adequate rules or definitions to capture these patterns, they soon realize they cannot, and, despite disagreement, are quick to acknowledge at least the cogency of each others’ cases and considerations. Whether or not they call them “analytic,” or are even justified in thinking in these ways, people do seem to have analytic intuitions, which I call “the analytic data,” leaving it open whether it is to be explained by an actual analytic.

Alas, however these analytic intuitions are a lot more slippery and elusive than the lovely examples of bad grammar. The history of philosophy is filled with purported analyses by one generation that have been refuted with counterexamples from the next:

\[\text{References}\]

See especially Katz (1972, pp. 3–6) for a rich survey of the data that seem to me to have been far from sufficiently discussed in the critical literature.
there is an entire sub-industry that has been devoted for some time to analyzing “x knows that p,” still without success [see, e.g., Morton (1997)].

Many of the difficulties in providing analyses, however, could be taken as actually an argument for, not against, the project of trying to find them, or something like them (it is a little like the combination of success and failure in science that argues for realism about its objects: we get enough success to think we are on to something, and enough failure to think that it is not all wishful thinking either). For what is it that leads us to reject proposed analyses? Why do we tend to agree about plausible candidates and also about their inadequacies? What do many of us know about “know” that leads us to fall for “justified true belief” and then to appreciate Gettier’s (1963) counterexamples? Or, what do we know about gods and ghosts to know that they had better have minds? Or, what do we know about [MURDER] and [PERSON] such that we understand why, for example, abortion and euthanasia are difficult cases? That is to say, we should ask here exactly the question that Chomsky (2000) asked about syntax: What explains the patterns and projections in peoples’ judgments?

An interesting case in point is Fodor’s (1981) rejection of any analysis of “paint v” (the transitive verb). He proposes and rejects a number of plausible proposed analyses, ending with:

(P) x paint v y iff x is an agent and x intentionally covers the surface of y with paint v and x’s primary intention in covering the surface of y with paint is that the surface of y should be covered with paint v in consequence of x’s having so acted upon it.

—Fodor (1981, p. 287)

Fodor (1981, p. 288) then raises still a further counterexample of Michelangelo dipping his brush into the paint, and so satisfying (P), but presumably not thereby “painting his brush.” At this point, Fodor concludes:

I don’t know where we go from here. For all I know – for all anybody knows – ‘paint v’ is undefinable; for all I know, you can’t eliminate it even in terms of such a closely related term as ‘paint n’. Or perhaps it is definable, but only in a reduction base that includes ‘dinosaur’ and ‘Chlorodent’. Either way, the present point is that [standard examples] don’t work. That’s not surprising; when it comes to definitions, the examples always don’t work.

—Fodor (1981, p. 288, emphasis his)

Now, first of all, it seems to me odd that Fodor breaks things off at this point – just when things are getting interesting! Several lines of continuation suggest themselves – specifying a still more detailed intention, restricting the scope of an analysis and allowing for

---

14 These are examples against the classical proposal that [KNOWLEDGE] is [JUSTIFIED TRUE BELIEF]; see Gettier (1963). Nichols, Stich and Weinberg (2003) present experimental evidence that people do not, in fact, converge in their reactions to these examples cross-culturally. However, as they acknowledge in response to Frank Jackson (1998), this could be because they are deploying different concepts – just as divergence about grammatical intuitions may be due to idiolectic variation. Indeed, given the role I have emphasized of concepts in guiding modal judgments, this would seem the reasonable conclusion to draw (once one controls also for the host of pragmatic influences surely also at work in the interview contexts). The appeal to intuitions is only as data to be explained, even if the data are diverse.
pragmatic variation in how “paint,” is understood in a context, perhaps understanding analyses as constitutive default rules, to be overridden in certain circumstances. But this is not the place to take up any of these suggestions. My points are methodological:

(i) How and why do we so readily see the plausibility of (P) and the cogency of Fodor’s further example? Is it really because (à la Quine) we have such a good empirical theory of the activity of painting? Could this be refuted by discovering that 6 out of 10 painters do paint their brushes? Does anyone care whether Michelangelo was painting his brush — apart from their grasp of the meaning of paint,? and

(ii) Why do we think uncovering the analysis of an item — determining either the primitives or the analysis — ought to be easy? Why do we not presume it will be at least as difficult as uncovering the principles of grammar?

In any case, being undefinable is hardly the natural conclusion to draw from the series of increasingly plausible analyses Fodor considered. We do seem to be converging on something, but, unsurprisingly, are maybe missing some subtle resources.

5.3. Rivals to the Analytic Explanation

One way to appreciate the role of some kind of analytic is to consider the alternatives.

5.3.1. Quine

One of the earliest and most influential alternatives was Quine’s own explanation, according to which what seems analytic is what is centrally or tenaciously believed by someone, capturing the idea that “no amount of evidence could refute the claim that bachelors are unmarried.” I will not rehearse all the inadequacies of this account. Suffice it to say that many unrevisable beliefs (e.g., “The world has existed for more than ten minutes.”; “There are many people in the world.”) are not the least analytic, and many analytic claims (e.g., “Bachelors are not married.”; “Kissing involves touching.”, “A podiatrist is concerned with feet.”) seem not the least “central”: indeed, the

15 This is no ad hoc maneuver. A number of philosophers and linguists have pointed to the immense difficulty of distinguishing semantic from pragmatic issues in evaluation of ordinary intuitions of “meaning”; see Bach (2001) and Pietroski (2003) for a discussion. For default rules, see Thomason (2001).

16 Consider, for example, how analyses of material object concepts might be fruitfully informed by the geometric constructions of vision theory, as in Marr (1982), Biederman (1987), and Palmer (1999), rather than by traditional sense data. And consider the considerable resources of Ramsey sentences in framing definitions in terms of (some selection of) the claims of a theory introducing and relating many terms at once [see Lewis (1972)]. The opera is not over until a good many people have sung a lot more arias.

Fodor worries in his discussion of “paint,” about how “it is vastly unlikely that children have access to [the concept] of “primary intention of an act” and that “the amount of subsidiary apparatus you need to define [the term] is getting sort of hairy” (1981, pp. 287–288). But it is hard to see why these should be worries, especially for someone, like Fodor who (in the same article!) is prepared to credit neonates with a fairly hairy innate conceptual endowment, especially with regard to language.

only reason most of us would resist giving up “bachelors are unmarried” is that *that is simply what the word means* – if someone wants to change the meaning to include women, fine! What is especially striking and interesting about purported analytic claims is not that they are *unrevisable*, but that, as *literally understood* their denials seem *unintelligible*.

### 5.3.2. Fodor

Fodor (1998) recognizes that Quine’s “centrality” explanation will not quite suffice to explain the analytic intuitions, and so follows up a suggestion originally advanced by Hilary Putnam (1962/1975): “analytic truths” are just examples of “one-criterion” concepts, or concepts like, e.g., [BACHELOR], [WIDOW], or [OPHTHALMOLOGIST], where there purportedly happens to be only one “way to tell” whether they apply. Fodor (1998, p. 82) acknowledges that, so stated, this latter account will not suffice either, since the notion of “criteria” seems no better off than “analytic.” In particular, if there were just *one* way to tell what’s what, there would seem, trivially, to be indefinite numbers of different ways – for example, just ask someone who knows the one way, or, ask someone who knows someone who knows the one way, etc. – and we would seem to have no better way to single out which way is “criterial” or which way is “analytic.”

But here Fodor (1998) offers a promising move, redolent of his general “asymmetric dependency” theory of content (see footnote 5): among the ways of telling “what’s what,” some do and some do not asymmetrically depend upon the others; the one(s?) that *do not* are the ones that give rise to analytic intuitions. Thus, telling that someone is a bachelor by checking out his gender and marital status does not depend upon telling by asking his friends, but telling by asking his friends does depend upon telling by his gender and marital status; and so we have an explanation of why “bachelors are unmarried males” seems analytic – without it actually being so.

Such asymmetric dependence will “explain away” the analytic, however, only if the analytic is not needed in an explanation of such asymmetric dependence. But it is hard to see how it would not be. Notice that not just any such asymmetric dependencies give rise to analytic intuitions: the only way I know to tell whether something is an acid is to see whether it turns litmus paper red; other ways (asking my chemist friends) asymmetrically depend upon that way. But “acids turn litmus paper red” is surely not analytic, even for me, since I bet there are, or could be, better ways to test acidity; and

---

18 At any rate, Fodor (1998) notes it does not work “for all the cases,” suggesting it works for some. I actually do not find it works for many cases, except as a matter of pure coincidence, e.g., cases where the term involved just happens also to be central to our thought, as in “Material objects are extended.”

19 Actually, one might wonder why Fodor’s (1998) proposal is not a reductive explication rather than the eliminative “explaining away” of the analytic that he intends; see also Horwich (1998), who claims meanings just *are* the uses of a term that explain all other uses. One reason might be the very reason I will now suggest against its even being a successful eliminativist alternative.
that is because “acid” marks a relatively “deep” sort of thing, whose nature is not captured by what tests happen to be at hand. By contrast, “bachelor” marks a superficial kind, whose nature is pretty much exhausted by the linguistics of the matter: unlike the case of litmus paper and acidity, the reason that gender and marriage status are the best way to tell whether someone’s a bachelor is that that is just what “bachelor” means! Indeed, should a chemist propose revising the test for acids in the light of better theory – perhaps reversing the dependency of certain tests – this would not per se constitute a change in the meaning. Should, however, a feminist propose, in the light of better politics, revising bachelor to include women, this obviously would.

In any case, it should be clear that appeals to “one criterion” concepts will not answer the explanatory challenge one might raise for the less obvious cases of the connections between, for example, [KNOW] and [JUSTIFIED], or between [FREEDOM] and [SPONTANEITY], since in each pair of concepts, although we seem to be on to some connections, it is notoriously difficult to specify even one genuinely adequate “way to tell”.

6. Conclusion

To summarize, reflection on our capacity to entertain primitive concepts with (necessarily) no referents, forces us to recognize that, pace strong externalists like Fodor and Millikan, but modulo Weak Externalism, some conceptual contents are individuated by some aspect of their role in our thought. Reflection on the variability of the roles of a concept, across people and within a single person across time, forces us to recognize that these roles are not necessarily available in a person’s superficial behavior or thought, but must be a part of an underlying competence. A traditional task of “philosophical analysis” could be regarded as aimed at uncovering those underlying roles. It is this task that seems to me of a piece with the task of an adequate psychology of concepts, which may, indeed, be as difficult as isolating the principles of a Chomskian grammar.

20 There is also the problem of how to apply the proposal to many of the obviously empty cases, such as [ANGEL] or [DEMON]: how, after all, does one actually tell whether something is an angel or the devil, as opposed to, for example, some misbegotten biology? Or whether someone has an immortal soul? Beats me. But I do presume that it is analytic that angels are some sort of supernatural mental agents, and that souls are the repository of our personal identity. Again, we look not to (how to determine) any angelic properties in the world, but merely to how people think about such other-worldly matters.
References


Rey, G. (Forthcoming a) Explanation, not experience. Commentary on John Campbell’s Reference and Consciousness, Philosophical Studies.
Chapter 4

CATEGORIES AND COGNITIVE ANTHROPOLOGY*

JAMES S. BOSTER

University of Connecticut

Contents

Abstract 92
1. Introduction 93
2. Cognition and culture, universalism and relativism 93
3. Paradigms and taxonomies 94
4. Kinship terminologies 101
5. Color classification 105
6. Ethnobiology 109
7. Towards a science of the stimulus 114
References 116

* My understanding of these issues has been greatly influenced by Roy D’Andrade, to whom I am very grateful. I would also like to thank him, Cornelia Dayton, Paul Kay, Asifa Majid, and Kateryna Maltseva for their comments on this chapter. I am grateful to David Kenny and John Gatewood for their help in locating Campbell’s writing on “entitity.” I thank Claire Lefebvre and Henri Cohen for their hard work in compiling the volume and Zofia Laubitz for her assiduous copyediting.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved

91
Abstract

In this chapter, I review how choices of methods of data collection and analysis have affected the representations of category structure by cognitive anthropologists in three domains: kinship terminologies, color classification, and ethnobiology. I argue that cognitive anthropologists have contributed to cognitive science in part by developing a “science of the stimulus,” leading to a better understanding of how the nature of the objects of experience affects the ways humans categorize them. I show that this approach offers an explanation of why some forms of semantic organization of lexicons are rare (e.g., paradigms) while others are common (e.g., trees and taxonomies), in part by expanding Campbell’s (1958) criteria for entitity to include both interdependence and discreteness of features. Domains whose features are independent of each other and vary continuously lack the entitity, in most cases, to be recognized as a class of things worth classifying. This observation, plus a comparison of the lexical overhead of naming unique combinations of feature values rather than the feature values themselves, helps explain the rarity of paradigms. Trees and taxonomies, conversely, are much more common because the domains they lexicalize have much greater entitity by these amended criteria. I discuss other examples of important advances in the cognitive anthropological understanding of particular domains that occurred when the methods and assumptions of the analysis better matched the nature of the domain being studied. In general, a “science of the stimulus” prompts the investigator to ask, “Where does domain structure come from?” In the case of ethnobiology, both the correlational structure of the domain and the cognitive architecture that perceives and categorizes it are crafted by the process of evolution by natural selection.
1. Introduction

This chapter reviews some of the contributions cognitive anthropologists have made to the understanding of human categorization. Rather than duplicate Roy D’Andrade’s (1995) overview of the history of the development of the field, I will focus on two questions. First, how do choices of methods of data collection and analysis affect the representations of category structure in various domains? Second, how do cognitive anthropologists contribute to a “science of the stimulus?”

The chapter begins by articulating the difference between cognitive anthropology and the prototype of the cognitive sciences – cognitive psychology – and presenting some historical background. This is followed by an examination of a portion of what cognitive anthropologists have learned in three domains: kinship, color, and ethnobiology. Although this review leaves out much more of the field than it includes, it will be used to answer the questions posed above.

2. Cognition and culture, universalism and relativism

If cognitive psychology is the study of the process of thought in individuals, observed in experimental settings, cognitive anthropology is the study of the content of thought (or knowledge) in communities of individuals observed in natural settings. Cognitive anthropologists seek to understand how culture happens, to explore how collective understandings of the world emerge in social groups, and to discover the pattern of cross-cultural similarities and differences in culture and cognition. The field sits at the intersection of psychology and anthropology, bridging studies of individual and collective representations.

The creation myth of cognitive anthropology might start, “In the beginning there was Murdoch.” After the Second World War, a very impressive collection of graduate students and junior faculty members gathered at Yale under the nominal dominion of George Peter Murdoch. Many of them, like Roberts, Conklin, and Goodenough, had served in the military in the war and were far more mature and toughened than the typical entering graduate students or assistant professors of other eras. Murdoch’s enterprise of cross-cultural comparison was well under way, producing Social Structure in 1949 (comparing 250 societies) and the “World Ethnographic Sample” in 1957 (comparing 564 societies). Although Murdoch himself had an encyclopedic knowledge of world cultures and devoured ethnographies at a pace that lesser souls might expend on

---

1 A discussion of cross-cultural variation in emotion categories can be found in another chapter of this volume.

2 Of course, there are many exceptions to this generalization – such as cognitive psychologists who do fieldwork and cognitive anthropologists who perform experiments – but the contrast as stated captures the difference in the central tendencies of the two disciplines.

3 I have heard various versions of this creation myth from Brent Berlin, Hal Conklin, Roy D’Andrade, and Jack Roberts. Whatever errors there are in this retelling are my own.
trash novels, he left the drudge work of coding the societies for the purposes of the ethnographic atlas largely to his graduate students. Jack Roberts recalled sitting in the library beside a tall stack of ethnographies, laboriously deciding what kind of wall construction was used for the houses, what kind of marriage and kinship system the society had, etc. The talent and intelligence of the graduate student workers combined with the immensity and drudgery of the task (and perhaps some elements of Murdoch’s personality) both grounded the students in a deep understanding of vast chunks of ethnographic literature and created the circumstances ripe for rebellion.

The rebellion came in the form of a sharp critique of the categories used to code the societies. The graduate students (by 2005, they are mainly distinguished professors emeriti) felt they were often forced to assign codes to societies in ways that did little justice to how the members of the society understood themselves. They wanted to develop methods that would allow the native’s point of view to emerge intact, without subjecting the description of a society to the procrustean violence of artificial coding categories. The rebellion produced something that would seem very strange in later decades: a fierce devotion to the possibility of extreme cultural relativism coupled with as fierce a faith in the power of the right formal methods of systematic data collection and analysis to deliver the native models intact. Thus, even in their rebellion, these early ethnoscientists bore the imprint of their association with Murdoch. At the same time that they rejected many of the assumptions that underlay Murdoch’s comparative enterprise, they embraced the rigor that characterized his systematic approach to cultural materials.

A further irony was that the discipline they started, which initially had been so committed to the possibility of extreme cultural relativism, found not only cultural differences but cultural universals as well. The tension between universalism (the notion that humans in different social groups, regardless of their natural, social, or linguistic environments, understand the world in essentially similar ways) and relativism (the notion that humans in different social and linguistic groups understand the world in very different, even incommensurate, ways) has created a great deal of debate in anthropology and, in the process, has probably wasted a great deal of ink and paper. Obscured by the debate is the fact that documentation of either position requires careful systematic data collection, analysis, and comparison. In developing the tools that could recover the native models intact in their distinctness, the early cognitive anthropologists also created the means to show that some native models were not limited to any one particular group of natives.

3. Paradigms and taxonomies

If one were to assign a date to the birth of cognitive anthropology, it would be the day Ward Goodenough delivered the paper that contained his classic definition of culture [Goodenough 1957 (1964, p. 36)]:

A society’s culture consists of whatever it is one has to know or believe in order to operate in a manner acceptable to its members, and to do so in any role that they accept for any one of themselves. Culture, being what people have to learn as distinct from their biological heritage, must consist of the
end product of learning: knowledge, in a most general, if relative, sense of the term. By this definition, we should note that culture is not a material phenomenon; it does not consist of things, people, behavior, or emotions. It is rather the organization of these things. It is the forms of things that people have in mind, their models for perceiving, relating, and otherwise interpreting them.

Some aspects of this definition were fairly standard. Most anthropologists would have agreed that culture includes knowledge and belief and that culture is learned. However, Goodenough’s insistence that both material artifacts and behavior in and of themselves were not cultural was unusual, and his assertion that culture consists of the models people use to understand the world and to guide their own behavior was novel. Also novel (for anthropologists) was the fact that the definition was an operational one. It outlined what a cultural description should look like and described a test for the adequacy of that description. An ethnography, according to Goodenough, should present a theory of the conceptual models people use rather than just a simple description of their behavior. It should be like a grammar or rule book for acting like a native. One tested the adequacy of an ethnography in much the same way that Turing thought we should test for artificial intelligence. If using the ethnography allows us to pass as a native, it is good enough.

Goodenough’s definition gave a very wide field for exploration – “whatever it is one has to know or believe …” – but the early cognitive anthropologists generally limited

4 One of the impacts of Goodenough’s definition was to show that how one defined culture could be a tool for articulating what one thought was important about human understandings of the world and how one thought it should be studied. He joined a minor cottage industry among anthropologists, each of whom had been defining culture in their own way. However, I interpret his contribution as saying, in effect, not that culture must be understood in his way, but rather that if one uses his definition, one will find out about a certain kind of aspect of how humans understand the world – the model-constructing, rule-seeking, grammar-building part. His contribution served as an invitation to other anthropologists to define culture in any way they find useful, as long as it bears enough resemblance to other definitions of culture to ensure that one is talking about roughly the same sort of thing. Definitions of culture by later cognitive anthropologists have tended to bear a family resemblance to Goodenough’s and to share with Goodenough’s the property of articulating what they think a cultural description should be about. For example, here is my own definition:

Culture is an information pool that emerges when members of a community attempt to make sense of the world and each other as they struggle and collaborate with each other to get what they want and need (e.g., food, sex, power, acceptance, etc.). Because individuals construct their conceptions of the world from their own experiences and for their own motivations, their understandings vary from one another depending on the characteristics of the individuals, the nature of the domain learned, and the social situations in which learning takes place. [modified from (Boster 1986)]

This definition shares Goodenough’s insistence that culture is learned knowledge, but adds that culture is (usually) useful and that individuals vary in their knowledge, and emphasizes that an ethnography should describe and explain the social distribution of knowledge.

5 The field was generally called “ethnosience” at the time and its practitioners “ethnoscientists.” The term “cognitive anthropology” was not coined until the publication of Tyler’s (1969) anthology. Ron Rohner, a fellow graduate student at Stanford in the early 1960s, takes credit for suggesting the title to Tyler.
their task to an exploration of the cognitive organization of lexicons. Frake (1962, p.74) offered a succinct justification for this narrowing of scope:

To the extent that cognitive coding tends to be linguistic and tends to be efficient, the study of the referential use of standard, readily elicitable linguistic responses – or terms – should provide a fruitful beginning point for mapping a cognitive system. And with verbal behavior we know how to begin.

The principal tool used to study the organization of terms in a language was componential analysis. This can be defined as a procedure for discovering the dimensions of meaning underlying a domain and mapping values on these dimensions (or components) onto terms in a lexicon. The procedure was directly analogous to those which had proved so useful to linguists in discovering phonemes and morphemes; one finds pairs or sets of terms that contrast with each other (i.e., are included in the same superordinate category), one attributes everything that is shared in meaning to membership in the superordinate category, and finally one discovers the attributes that distinguish among the members of the contrast set. While Frake’s (1962) description of the method allowed it to be applied to taxonomies (nested contrast sets related by inclusion), Goodenough (1956) and Lounsbury (1956) further restricted the application of the method to those lexicons that formed paradigms, in which every (or nearly every) combination of feature values corresponded to a distinct term.

The most perfect example of the application of componential analysis was Conklin’s (1962) analysis of Hanunóo pronouns shown in Figure 1. The eight pronouns corresponded to every possible combination of three binary feature contrasts – inclusion and exclusion of speaker (S, S), inclusion and exclusion of hearer (H, H), and minimal and non-minimal membership (M, M) – and the whole set of pronouns could be represented as the vertices of a cube, whose depth, width, and height corresponded to the three feature dimensions.

Unfortunately, examples of such heart-squeezing beauty and elegance were rare. Beyond a few applications of componential analysis to pronominal systems and kinship systems, there were few domains organized as paradigms. And even kin terminologies fit awkwardly into strict paradigms. One reason for the rarity of paradigms can be seen by comparing the number of terms required to label every possible combination of feature values with the number required to label the feature values themselves, as shown in Table 1.

---

6 The property of language exploited in these methods was its capacity to make near infinite use of finite means. At every level of the analysis of language (phonology, morphology, semantics, and syntax), one finds a finite set of elements (phonemes, morphemes, grammatical classes) combined through a finite set of rules (phonological rules, morphological rules, syntactic rules) to generate a much greater number of possible forms. Linguists (and cognitive anthropologists) use this property of language to work backwards, systematically comparing minimally different complex units (sounds, words, meanings) to discover the underlying universe of elements and the rules for combining them.

7 This use of “paradigm” is in the same sense as “verb paradigm,” the pattern of verb inflections which often take different forms according to whether speaker and/or hearer, etc., are included in reference, etc. It has nothing to do with Kuhn’s (1962) use of the word “paradigm” in the sense of a scientific framework.
If $N$ is the number of binary feature dimensions, the number of terms required to label the unique combinations of features is $2^N$, while the number of terms required to label the distinct feature values is $2N$. Thus, as the number of feature dimensions increases, the lexical cost of labeling the terminal nodes of a perfect paradigm relative to labeling the feature values (the ratio $2^N/2N$) increases exponentially. A feature key of a perfect paradigm [using the conventions of Kay (1966)] is shown in Figure 2.

Feature dimensions are indicated by letters (e.g., “a,” “b,” “c,” …) and feature values are indicated by subscripts (e.g., “a$_1$,” “a$_2$,” “b$_1$,” “b$_2$,” “c$_1$,” “c$_2$,” …). (For example, “a” could be used to signify the feature dimension of sex, and “a$_1$,” “a$_2$” used to signify the feature values of male and female, respectively.) With two dimensions, one needs just as many terms (four) to label the unique combinations of features as to label...
the distinct feature values. With three dimensions, one needs a third again as many terms to label the unique combinations as to label the feature values, with four dimensions, twice as many, and by the time one gets to six feature dimensions, it requires more than five times as many terms to label the unique combinations of feature values as to label the unique features alone. Six feature dimensions is the theoretical limit according to Wallace (1964), corresponding to Miller’s (1956) “seven plus or minus two” constraint on short-term memory. If it were not for this constraint, the disparity would be 2,048 times as many terms with 16 feature dimensions, and so on. No wonder paradigms are rare; very frequent reference to the unique combinations of feature values (as in pronominal systems) is required to justify the lexical overhead.

However, a more profound reason for the rarity of paradigmatic organization of lexicons may be that the very independence of feature dimensions that would allow them to be represented in neat nested boxes deprives them of the “entitivity” necessary to be recognized as classes of things that require their own labels. “Entitivity” is a word coined and defined by Campbell (1958, p.17) as “the degree of being entitative. The degree of having the nature of an entity, of having real existence.” Campbell was addressing the question of when one calls aggregates of human beings a social group. In order to do so, the group had to have some degree of entitivity, so as to distinguish accidental agglomerations of people (e.g., the passengers waiting at a bus stop) from members of more “real” groups (e.g., the lifetime members of a professional organization, the members of a nuclear family). Campbell listed five criteria: common fate, similarity, proximity, resistance to intrusion, and internal diffusion. Campbell was building on insights by Herbert Spencer (1897), who regarded a living body as an entity par excellence. Most of Campbell’s criteria were borrowed from those that Gestalt psychologists used to decide whether something constituted a gestalt. I would add two more criteria to his list when deciding whether a collection of things constitutes a set of entities worth classifying: “do the features that characterize the items vary independently or interdependently?” and “do the features vary continuously or are the feature values discrete?” Both interdependence of feature correlation and discreteness of feature values contribute to the “entitivity” of items in a domain.

---

Fig. 2. Feature key of a perfect paradigm. Adapted from Figure 2a, of Kay (1996).
Consider soils. They vary continuously in texture from very fine clays through silts and loams to coarse sands and gravels. The variation in texture is largely independent of the equally continuous variation in color, as indicated by the huge number and subtle gradations of colors found in the Munsell soil color book. If every single possible combination of color and texture of soil were categorized as a distinct kind of soil and given its own label, our folk soil classification would dwarf our folk biological classification. This does not happen. Folk systems tend to have a few terms like “clay,” “silt,” “loam,” and “sand” that describe variation in texture and a few other terms like “muddy,” “parched,” and “rocky” to describe degrees of saturation with water or stones. Folk systems have the possibility of describing the colors of soils just as they can describe the color of almost anything else. However, they do not have nearly the number of terms necessary to distinctly name all possible combinations of variation in texture and color. The K’ekchi Maya corn farmers I worked with some years ago classified soils principally into just two classes: *chabil ch’och* (‘good earth’) and *ma us aj’ ch’och* (‘bad earth’), distinguished by their suitability for corn farming. Even soil scientists rarely use the named classification of soils, instead describing the horizons of the soil profile by their color (determined by consulting the Munsell soil book) and texture when surveying soils. Although there is sufficient variation in the characteristics of soils to allow many different types of soil to be differentiated, because of the continuity and independence of the feature dimensions, they do not make for “discontinuities in nature calling out to be named.” They also lack other aspects of entitivity described by Campbell, especially the lack of any clear boundaries between one soil type and another and the absence of any appreciable resistance to intrusion.

The greatest degree of entitivity, using Campbell’s amended criteria, is found among living things, as Spencer intuited. Biological evolution is a wonderful structure-generating engine. Each speciation event distinguishes a pair of species on at least one and usually several features. Once they are no longer part of the same breeding population, the two new species may respond to different sorts of selection pressures, so that the granddaughter species cleaved by subsequent speciation events are usually not distinguished on the same features as the other granddaughter species. The feature key (again, using the conventions of Kay, 1966) representing this process is shown in Figure 3, which not coincidentally, is also the feature key of a perfect tree.

In this case, there is a lexical economy in naming the unique combinations of feature values rather than the feature values themselves, because the features multiply faster than their unique combinations do, as shown in Table 2. Again, if *N* is the number of feature dimensions, the number of terms required to label the unique combinations of features is *N*+1, while the number of terms required to label the distinct feature values themselves is 2*N*. Thus, as the number of feature dimensions increases, the lexical cost of labeling the terminal nodes of a perfect tree relative to labeling the feature values (the ratio (*N*+1)/2*N*) decreases asymptotically to ½.

Another difference is that in a paradigm there is no intrinsic ordering of the feature dimensions, while in a tree there is an intrinsic order. Kay (1966) referred to this difference as the contrast between the simultaneous application of feature dimensions in a
paradigm versus the sequential application of feature dimensions in a tree. Because a particular feature dimension applies at one and only one node of a tree, labeling the superordinate nodes is meaningful in a tree in a way that it is not in a paradigm. If the superordinate nodes of a tree are labeled, it is called a taxonomy. A third difference between paradigms and trees is their difference in feature redundancy. The excess of distinguishing features over the number of categories to be distinguished in trees makes it much more likely that the combinations will be “chunked” into an integrated gestalt, something that the feature independence of paradigms does not afford. It is the interdependence of the features of trees that Rosch and Mervis (1975) are referring to with their phrase “correlational structure.” The greater lexical economy in labeling unique combinations of feature values, the greater entitvity of the items in the domains, and the more memorable “gestalts” of feature combinations make trees and taxonomies much more common modes of semantically organizing lexicons than paradigms. This meant that componential analysis was soon displaced by other methods of exploring the semantic organization of lexicons as the principal tools of cognitive anthropologists, especially by those based on taxonomies [e.g.,

Table 2

<table>
<thead>
<tr>
<th>Number of binary feature dimensions (N)</th>
<th>Number of terms to label to combinations of feature values (N + 1)</th>
<th>Number of terms needed to label feature values (2N)</th>
<th>Lexical cost of labeling combinations of feature values relative to labeling feature values directly ((N + 1)/2N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>6</td>
<td>0.67</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>14</td>
<td>0.57</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>30</td>
<td>0.53</td>
</tr>
<tr>
<td>31</td>
<td>32</td>
<td>62</td>
<td>0.52</td>
</tr>
<tr>
<td>63</td>
<td>64</td>
<td>126</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Fig. 3. Feature key of a perfect tree. Adapted from Figure 2b, of Kay (1996).
Metzger and Williams (1966)] or open-ended semantic networks [e.g., Frake (1964)] and by methods borrowed from psychology [e.g., Romney and D’Andrade (1964)].

4. Kinship terminologies

An early demonstration of the way in which the choice of data analysis method could affect the representations of category structure came in the analysis of kinship terminologies. As a way of showing off the power of their methods, various early cognitive anthropologists analyzed the structure of American English kin terms [e.g., Wallace and Atkins (1960), Romney and D’Andrade (1964), Goodenough (1965)]. As Schneider (1965) has pointed out, using American kinship as the arena for the comparison of methods allowed readers who were native speakers of American English to criticize the analyses with their own native intuitions, unlike their helplessness in the face of descriptions of the exotic. Two theoretical questions came to the fore in this exchange: “how do you choose between alternative analyses?” and “how do you demonstrate the psychological validity of a particular analysis?” Both questions stemmed from the sense that an analysis of a semantic domain should reflect the cognition of the informants, not the investigator.

Wallace and Atkins (1960) described the method of componential analysis of kinship terminologies propounded by Goodenough (1956) and Lounsbury (1956) as consisting of five steps: (1) recording a complete set (or defined subset) of kin terms; (2) defining the terms by their kin types (e.g., MoFa, MoBrSo); (3) identifying two or more feature dimensions whose values or components are signified by the terms; (4) defining each term by its feature values; and (5) stating the semantic structure of the whole set of kin terms. Starting with a defined subset of English kin terms (the consanguineal core terms), they end up with the structure shown in Table 3.

In their discussion, Wallace and Atkins recognize that alternate analyses are possible and that one should prefer the analysis that is “the most proximate to psychological reality” (p. 78). However, the authors recognize the difficulty of developing methods to establish psychological validity and do not actually present any themselves.

<table>
<thead>
<tr>
<th>lineal</th>
<th>colineal</th>
<th>ablineal</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>female</td>
<td>Male</td>
</tr>
<tr>
<td>+2</td>
<td>grandfather</td>
<td>grandmother</td>
</tr>
<tr>
<td>+1</td>
<td>father</td>
<td>father</td>
</tr>
<tr>
<td>0</td>
<td>[ego]</td>
<td>daughter</td>
</tr>
<tr>
<td>−1</td>
<td>son</td>
<td>daughter</td>
</tr>
<tr>
<td>−2</td>
<td>grandson</td>
<td>granddaughter</td>
</tr>
</tbody>
</table>

Adapted from Figure 1 of Wallace and Atkins (1962).
Goodenough (1965) later analyzed a greater range of kin terms (including step- and in-law terms) and solved the problem of psychological validity to his own satisfaction by using himself as his own informant and rejecting analyses that “just didn’t sit right.” His methods and resulting analysis, however, are substantially similar to those of Wallace and Atkins (1960).

In contrast, Romney and D’Andrade (1964) present a radically different approach both to the problem of producing an analysis and to that of establishing its psychological validity. They begin with the same set of consanguineal core kin terms used by Wallace and Atkins (1960), but define the terms using a special notation to indicate the kin paths between ego and various alters, rather than the usual kin type notation. In this notation, “m” indicates a male; “f” a female; “a” someone of either sex; “0” a sibling link; “+” a parent link; “−” a child link; “=” a marriage link; and “/” is used to enclose expressions that can be read as reciprocals. Thus, a mother’s brother’s son would be represented as “a+fom” rather than “MoBrSo.” The advantage of this notation is that it allows the authors to describe a set of rules that reduce to a single expression all the paths linking ego to various alters indicated by a kin term. Romney and D’Andrade’s (1964) procedure produces the structure shown in Table 4.

Note that the principal differences from Wallace and Atkins (1960) are that there are only two degrees of lineality (direct and collateral) rather than three (lineal, colineal, and ablineal) and that generational terms that are reciprocals of one another (e.g., grandmother, granddaughter; mother, daughter) are treated as sharing the feature value of being one or two generations away from ego, rather than distinguishing kin terms by their absolute generation.

Romney and D’Andrade’s (1964) second major innovation was to describe a set of procedures to compare the psychological validity of their own model with that of Wallace and Atkins (1960). The critical data for the comparison were elicited using a triad test in which they presented all possible triplets of a set of eight male kin terms to 116 high school students and asked of each triad, “Which term is most different from

| Table 4 |
|---|---|---|---|
| **direct** | **collateral** |
| | male | female | male | female |
| two generations | + | grandfather | grandmother | male | female |
| different | − | grandson | granddaughter |
| one generation | + | father | mother | uncle | aunt |
| different | − | son | daughter | nephew | niece |
| same generation | | brother | Sister | cousin | |

Adapted from Figure 3 of Romney and D’Andrade (1964).

Same-generation terms are reciprocals of themselves (e.g., I am my brothers’ brother and my cousins’ cousin.)
the other two?” They found that terms that were reciprocals of one another (e.g., grandson and grandfather) were judged as more similar than were terms that were closer in absolute number of generations (e.g., grandfather and father), as predicted by their own model and not by Wallace and Atkins (1960). Furthermore, the three dimensions of a multidimensional scaling of the similarity judgments corresponded to the three dimensions of Romney and D’Andrade’s (1964) model: reciprocity, generational difference, and lineality. They concluded that their model has greater psychological validity than the alternative model described by Wallace and Atkins (1960).

Little turns on this particular conclusion, for it hangs on a slender thread. True, when given the triad grandfather, father, and grandson, most American English speakers will judge father to be the most different from the other two. However, a subtle change in wording produces a contrary result; if given the triad my grandfather, my father, and my grandson, most will pick my grandson as the most different, as predicted by Wallace and Atkins (1960). Notice that the change in wording has this effect because it shifts attention away from the relationship ‘grandfather’ or ‘father’ to a particular individual who is someone’s grandfather or father. The mature elders who are fathers and grandfathers are much more similar to each other than the callow youth who is their son or grandson, even if the relationship ‘grandfather’ is more similar to its reciprocal ‘grandson’ than is to the relationship ‘father’ or ‘son.’ This shift of attention from relationships to individuals better matches the typological assumptions of componential analysis.

More significant is Romney and D’Andrade’s development of methods for establishing the psychological validity of alternative models of semantic structure. But here again, although this was an important step forward at the time, in retrospect, it has had little lasting effect on the field, because it was not followed by a flood of papers that developed a model using one set of methods and tested its validity with another set. Instead, the psychological and statistical methods that Romney and D’Andrade (1964) introduced to test for psychological validity soon displaced the primarily linguistic methods that had hitherto been used to develop the models. Investigators began eliciting similarity judgments to directly build their models of semantic structure and dispensed with componential analysis and related methods altogether. Thus, multidimensional scalings and cluster diagrams of similarity judgments replaced the keys and boxes of taxonomies and paradigms as figures in cognitive anthropological articles. In the process, the notion of testing models of semantic structure for their psychological validity with an independently collected set of data was lost.

There is also the sad fact that interest in studying kinship and kin terminologies has largely withered among anthropologists and linguists. Hardly anyone noticed a few years later when Ellen Woolford (1984) offered the definitive analysis of kinship terminological systems. Woolford combined elements of three different approaches (componential

9 By “typological assumptions,” I mean the assumptions that the terms in a lexicon label a set of items in a semantic domain and that each of the items can be distinguished from the others by some small set of distinguishing features. In the case of kinship terminologies, the terms refer to relations between items (the individual humans); they are not distinguished by features of the items in isolation.
analysis, extension rule analysis, and relational analysis) in a clear and elegant generative model that predicts a number of implicational universals (e.g., “if you refer to the same sex siblings of your parents with the same terms you use for your parents, you will refer to their children with the same terms you use for your siblings”). The field has not died altogether. Dwight Read and his collaborators continue to analyze kin terminologies with (kinship analysis expert system) (KAES) [(Read and Behrens, 1990)]. However, one needs a test similar to that performed by Romney and D’Andrade (1964) to decide whether KAES produces more psychologically real models than its predecessors.

I believe that the most important contribution of Romney and D’Andrade’s (1964) work has been little appreciated. The reason why their procedure produced an analysis that was superior to that produced by componential analysis is that the algebraic analysis of relationship paths is truer to the logic of the domain of kinship than the typological assumptions underlying componential analysis. That is, although the term niece picks out of the world a small set of particular (and charming) human beings for me, the term really does not refer to particular humans but to a relationship between humans. I call them nieces, they call me uncle, but all of us bear a variety of kin relationships to others. In other words, I am not an ‘uncle’ in general, but only an ‘uncle’ to my nieces and nephews. The algebraic notation and reduction rules developed by Romney and D’Andrade (1964) capture the logic of these relationships better than any typological method. It is a shame that the authors demonstrated the approach by analyzing American English kin terms, for its real power is shown off best in making the exotic transparent, as D’Andrade (1995, pp.20–28) shows in analyzing Bellah’s (1952, p.71) data on Chiricahua kin terms10. Using Romney and D’Andrade’s (1964) notation and reduction rules, one can derive a single expression \( /a+m(+a)0(a-)a/ \) which represents the 40 relationship paths from ego to various alters that are referred to by the Chiricahua kin term \( \text{cìdèèdèè} \). (Similar reductions can be accomplished for the other Chiricahua kin terms.) Romney and D’Andrade (1964) heralded what became a common pattern in the development of cognitive anthropology. Our understanding of particular domains was often improved when the methods and assumptions of the analysis better matched the nature of the domain being studied11.

Another work of this period, Nerlove and Romney (1967), also had far-reaching implications, not so much for the importance of the domain studied (sibling terminologies) as for its demonstration that the methods that had been used to explore the internal logic of a lexicon in a single language could be used to systematically compare semantic systems. The first step was to describe a universe of logical possibilities. There are eight possible kin types given by all possible combinations of sex of ego (\( x = \text{male}, o = \text{female} \)), sex of alter (\( B = \text{male}, Z = \text{female} \)), and age difference of alter and ego (\( e = \text{elder}, y = \text{younger} \)) and there are one to eight possible kin terms to refer to these

10 Shortly after the publication of Romney and D’Andrade (1964), Romney (1965) applied the method to the analysis of Kalmuk Mongol kinship terminology but this article never became as well known as the earlier work.

11 John Atkins’ (1974) GRAFIK metalanguage was arguably a superior algebraic notation system for kinship.
kin types, yielding a potential universe of 4,140 possible ways of assigning kin types to categories labeled by the kin terms. (For example, in English, the kin types $xZe$, $oZe$, $xZy$, and $oZy$ are all called sister and the kin types $xB$, $oB$, $xB$, and $oB$ are all called brother. In contrast, in Awajun (Aguaruna Jivaro), the kin types $oZe$ and $oZy$ are called kai, the kin types $xB$ and $xB$ are called yatsu, and the kin types $xZe$, $xZy$, $oB$, and $oB$ are called uma.) Nerlove and Romney’s next step was to characterize the sibling terminologies found in a sample of 245 societies using this notation system.

Out of the 4,140 conceivable ways of categorizing the kin types, Nerlove and Romney (1967) found only 12 regularly occurring types with a handful of outliers in their sample. Thus, the cross-cultural variation in sibling terminologies is both much greater than would be guessed from English alone (only 10% of societies distinguish siblings according to absolute sex alone) and much smaller than the range of possible variation. Nerlove and Romney explained the constraints on the variation in terms of universal cognitive and social factors (e.g., disjunctive categories are rare compared with conjunctive or relational categories; categories in which sex of ego is a primary distinction are nonexistent), and they explained the variation in terms of social variables (e.g., sibling categories in which relative sex of ego and alter is a primary distinction are most common in societies that have brother–sister avoidance and long postpartum sex taboos). This article was an important development because it used explicit formal methods to explain patterns of cross-cultural similarities and differences. Although this work has been singled out by Shweder and Bourne (1984, pp.160–161) as exemplifying a mode of finding cultural universals by excessively circumscribing the scope of study, such circumscription to sibling terms, and not kinship terms in general, was probably necessary to be able to identify a coherent and tractable comparative project. Romney and D’Andrade (1964) also represented a rejection of the relativism of the early cognitive anthropologists and a return to Murdoch’s systematic cross-cultural comparison. This became the model for later cross-cultural comparison of categorization in color, ethnobiology, and other domains as discussed below.

Before we turn our attention to the development of an understanding of color classification by cognitive anthropologists, it is useful to review the steps used by Nerlove and Romney (1967) that were emulated in later investigations of patterns of cross-cultural similarities and differences. First, one picks a domain and examines the structure of the feature space. Second, one isolates comparable units of analysis. Third, one compares the systems and describes patterns of similarities and differences. Finally, one attempts to explain the similarities in terms of shared constraints, and the differences in terms of factors that differentially affect the societies. Unfortunately, cognitive anthropologists did not always get around to performing this last step.

5. Color classification

Color is another domain in which the choice of method (especially the choice of how the units of analysis were defined) affected the representation of category structure.
Shifts in theoretical interpretation were also driven by the collection and analysis of an ever-increasing quantity and quality of data. Early studies treated color classification as an example of extreme cultural relativism. In Roger Brown’s textbook, *Social Psychology*, the color spectrum was represented as a continuum that is segmented by different languages in arbitrarily different ways (1965, pp. 315–316). The empirical reports that were available supported the extreme relativist position. For example, Conklin’s (1955) study of Hanunóo color categories described them as encoding features including surface texture and sheen, not just hue, brightness, and saturation. Furthermore, one would expect cultural relativism in color categorization given the nature of the domain. Colors vary continuously from one another in hue, brightness, and saturation. The continuity and independence of the feature dimensions deprive the domain of entitivity. If there is no intrinsic structure to the color space to guide their category boundaries, nothing would prevent speakers of a different language from cutting the continuum up in a way that is arbitrarily different from the way English does it.

But this view was not at all correct. As Paul Kay in Tahiti and Brent Berlin in Tenejapa observed from their experiences doing fieldwork, it was not nearly as difficult to learn the meanings of the color terms in Tahitian or Tzeltal as would be implied by an extreme relativist position. They addressed the question in a perfectly sensible way for assistant professors at Berkeley: they organized a graduate seminar and sent their students throughout the San Francisco Bay area with Munsell color charts to see how the color terms in different languages mapped to a common set of referents. What they found won them fame and tenure, and established what they later called the Universals and Evolution approach to color classification. The breakthrough was achieved by defining color categories by their best examples, because, although there was considerable interinformant variation on the placement of the boundaries between color categories (even with the use of a stimulus array that encouraged well-bounded categories), there was considerably more agreement both within and between languages on the choices of the best examples, or foci, of the categories. Berlin and Kay (1969) found that there were only 11 basic color categories in the 20 languages they were able to investigate, with foci in black, white, red, yellow, green, blue, brown, gray, pink, orange, and purple. This confirmed their initial hunch. More surprisingly, the color categories came in a limited number of combinations. If there were two categories, their foci were in black and white; if there were three categories, they focused in black, white, and red; if four categories, black, white, red, and yellow, or black, white, red, and green; if five categories, black, white, red, yellow, and green; if six categories, blue was added; if seven, brown was added; and if eight or more categories, gray, pink, orange, and purple were added in no particular order. This pattern was interpreted as the sequential encoding of foci, as though the speakers of the various languages had a push-down stack of color foci (like the floating stacks of plates found in some cafeteria lines), with white and black at the top (at stage I), red next (stage II), yellow or green next (stage IIIa or IIIb), green or yellow next (stage IVa or IVb), followed by blue (stage V), brown (stage VI), and a random shuffle of gray, pink, orange, and purple foci at the bottom (stage VII). Berlin and Kay (1969) had to guess what a two-category system would look like...
like because none of the languages they had access to had only two terms, but their framework accommodated all of the systems that they and their students were either able to investigate directly or surmise from dictionary entries.

But this model was not quite right. When Eleanor Rosch (then Heider) later investigated a two-term color classification system among the Dani of Irian Jaya, she found that the ‘white’ category actually focused in white, yellow, or red and was better described as a warm/light category, while the ‘black’ category actually focused in black, green, or blue and was better described as a cool/dark category [Heider (1972a,b)]. Similarly, Berlin found that the ‘green’ term in Tzeltal Mayan and other stage IV systems could focus in either green or blue and was better described as a [GRUE] category. To accommodate these new data, Kay and McDaniel (1978) reinterpreted the evolutionary\(^{12}\) sequence (at least until stage V) as the successive differentiation of existing color categories. Thus, a two-term system at stage I was interpreted as having two categories of unique hue points: white, red, and yellow versus green, blue, and black. With three categories, white was split from red and yellow; with four categories, either yellow was split from red or green and blue from black; with five categories, red and yellow were split apart, and green and blue were split from black; and with six categories, all six of the unique hue points were split into separate categories. Secondary basic color terms were interpreted as the intersection of unique hue points: brown as the intersection of yellow and black; gray as the intersection of white and black; pink as the intersection of white and red; orange as the intersection of red and yellow; and purple as the intersection of red and blue. Languages could have categories defined by the intersection of unique hue points as soon as the constituent hue points were in their own separate categories. This helped account for the “premature gray” found in some Amerindian languages.

Kay and McDaniel (1978) also argued that the unique hue points and the cross-cultural universals they gave rise to had their origin in the neurophysiology of color vision. They based their interpretation on DeV alois, Abramov and Jacobs (1966) research on the lateral geniculate nuclei (LGN) in the hypothalamus of Rhesus macaques which had revealed three families of neurons. Two families were opponent processes: a red–green channel (excited by red light and inhibited by green light, or vice versa) and a yellow–blue channel (excited by yellow light and inhibited by blue light or vice versa). The third family was a white–black channel which responds to brightness levels independently of the other two channels. This physiological account helped explain why there might be universals in the classification of what seemed a structureless domain. The structure is imposed by the neurophysiology of color vision.

But this schema was still not quite right. One problem was that anomalous composite categories were found, which joined yellow, green, and blue; yellow and green; white and yellow; or blue and black. None of these combinations could be derived from the successive differentiation of color categories as described by Kay and McDaniel (1978). Furthermore, subsequent research had revealed that the neurophysiological opponent process system would actually put the unique hue points in the wrong places. The true

\(^{12}\) The term “evolutionary” in this context refers to lexical, not biological, evolution.
axes of the system are closer to cherry–teal and chartreuse–violet than they are to red–green and yellow–blue [Jameson and D’Andrade (1997)]. Sadly, this left the universals in color classification without a clear physiological explanation. Repairs to the framework were made by Kay, Berlin and Merrifield (1991), which provided the first report of the World Color Survey (WCS). The WCS investigated color naming in 111 languages, interviewing roughly 25 native speakers of each language about their color names for each of 330 color chips, and their choices of the best examples of each basic color term of their language. The WCS represented an enormous improvement in both the methods and the quantity of data over what was available to Berlin and Kay (1969). The 330 color chips were presented to informants one by one in random order, thereby overcoming a bias toward the recognition of contiguous color categories that was inherent in the use of the original Munsell grid of color chips arranged in 8 rows and 40 columns (plus the gray scale). Many more speakers of each language were interviewed, thereby allowing a genuine comparison of the magnitudes of intra- and interlanguage variation. And finally, nearly six times as many languages were surveyed, most of which were spoken by indigenous groups in Africa, Central and South America, Papua New Guinea, and smaller numbers of indigenous groups from North America, India, Indonesia, the Philippines, and Australia, all areas where the linguists of the Summer Institute of Linguistics were working. Most of the languages sampled had early stage color systems. Because so many more languages were studied beyond the initial set of 20 in Berlin and Kay (1969), it was practically inevitable that the researchers would encounter greater variation than they had initially. To account for this greater variation, they formulated the “composite category rule” which posited a network of unique hue points in which white, red, and green were each independently linked to yellow, blue was linked to green, and black linked to blue as shown in Figure 4. Their composite category rule allowed for the existence of any composite category that joined linked hue points which did not cross yellow. This rule allowed all of the composite categories permitted by Kay and McDaniel (1978) (red, white, yellow; green, blue, black; red, yellow; green, blue) plus a number of additional composite categories which had previously been treated as anomalous (yellow, green, blue; white, yellow; yellow, green; and blue, black). Their rule also allowed a composite category that had not yet been found (yellow, green, blue, black).

![Fig 4. Network of unique hue points in the composite category rule. Adapted from Figure 2 in Kay, et al. (1991).](image-url)
But even this formulation was not quite right. The problem was that the solution to the existence of anomalous categories was awkward; the composite category rule was an arbitrary statement of permitted categories and offered no statement of its own logic. It was a step backwards, and one that Kay et al. (1997) and Kay and Maffi (1999) rectified. The rectification was a partial reversion to the architecture of the original Berlin and Kay (1969) evolutionary sequence, but instead of recognizing just two evolutionary trajectories, they now recognized five. The five trajectories were interpreted as generated by four principles: partition (lexicons tend to partition items in culturally significant domains into exhaustive and mutually exclusive categories); black and white (color lexicons distinguish black and white); warm and cool (color lexicons distinguish the warm primaries from the cool primaries); and red (color lexicons distinguish red from other colors). Mysteriously, the yellow-white composite category found in Waorani, an Amazonian Amerindian language, was no longer a permitted composite category, but this was a small price to pay for the much greater clarity and coherence of the framework.

But this picture is unlikely to be quite right, given the long history of revisions of the framework. Cook, Kay and Regier (this volume) offer the latest results of the analysis of the WCS. Still, I believe this history is a heartening one. Although the most recent statement of the evolutionary sequence is not nearly as simple as the original one, it is based on much better evidence from about six times as many languages as the original. The revised framework manages to preserve the most important element of the Universals and Evolution tradition: the notion that underlying the considerable cultural variation in how languages categorize colors, there is nevertheless evidence of a universal pattern.

### 6. Ethnobiology

Ethnobiology, the study of the folk classification of plants and animals, is a third domain that illustrates the impact of changes in methods of data collection and analysis on our understanding of the nature and structure of semantic domains. In particular, two works by Berlin and colleagues illustrate how the shift from a relativist to a universalist interpretation came about mainly by virtue of a shift in how the units of analysis were defined. Berlin, Breedlove and Raven (1966, p. 273) clearly and simply defined a folk specific category as “any taxon which includes no other taxa.” This definition had the advantage of unambiguously specifying the units of analysis as the set of all terminal taxa – categories not further subdivided. They then examined the ways in which the folk specific taxa corresponded to scientific species. Forty-one percent of their sample of 200 Tzeltal folk specifics were underdifferentiated with respect to scientific species (a single Tzeltal folk specific included more than one scientific species), 34 percent were in a one-to-one correspondence with scientific species, and 25 percent were overdifferentiated with respect to scientific species. They concluded that “Tzeltal specifics clearly do not correspond in any predictable way with botanical species” (1966, p. 273).

Berlin et al. explained the variation in the way folk specifics map to scientific species by appealing to the cultural significance of the plants. Culturally significant plants such
as food crops and other useful plants were likely to be overdifferentiated, economically unimportant species were likely to be underdifferentiated, and introduced species were more likely to be in one-to-one correspondence. Furthermore, those folk specifics that are overdifferentiated with respect to scientific species often had the form of “attributive + head” where the head term referred to a category in one-to-one correspondence with a scientific species (e.g., terms like *sweet corn, pop corn, dent corn; corn = Zea mays*).

A short time later, Berlin et al. had a change of heart when they looked at their data in a different way. This is an unusual case [as noted by Gould (1980)] in which the same team of researchers investigating the same people’s knowledge of the plant world completely reversed themselves, and shifted from a relativist to a universalist interpretation of essentially the same data. The critical shift came in how they defined a folk species. Berlin, Breedlove and Raven (1973, p. 214) state this change of heart:

The richness and diversity of man’s variant classifications of experience have often led ethnographers to emphasize the differences between cultural systems of knowledge…. There are a number of strikingly regular structural principles of biological classification which are quite general.

Their principles were that ethnobiological domains in all languages are organized taxonomically; that taxa are grouped into a small number of ranks (unique beginner, life-form, generic, specific, varietal) that are hierarchically arranged; taxa of the same rank usually occur at same level of the taxonomy and are mutually exclusive. The most inclusive rank, the unique beginner, is often unlabeled; life-forms are immediately included in the unique beginner, are few in number (5–10), include most lower taxa, and are polytypic (e.g., *tree, vine, bird*). Categories at the folk generic rank are the most numerous, although most are included in life-forms, some are unaffiliated, but all are highly salient and commonly referred to, learned early by children, and labeled by primary lexemes (e.g., *oak, pine, robin, cactus, banana*). Specific and varietal categories are fewer in number, occur in small contrast sets, are often culturally important, distinguished by few features, and labeled by secondary lexemes (e.g., *live oak, valley oak, Ponderosa pine, Monterey pine, butter lima bean, baby lima bean*). Finally, intermediate categories are included in life-form categories and include generics, are often unlabeled, and are very rare (e.g., *evergreen, deciduous, songbird*). The advantage of these principles of folk biological classification is that they allowed one to recognize a much greater degree of correspondence between folk and scientific systems of classification, and also much higher degrees of structural similarity among the folk systems themselves. For example, the “attributive plus head” term found among many overdifferentiated folk specifics in the previous (1966) scheme, remained folk specifics in the new system, but the head term without any attributive was recognized as a folk generic, and thus more comparable to the categories that were in one-to-one correspondence or underdifferentiated in the earlier scheme. The cost of this insight was much greater overhead in complicated rules to distinguish between categories at the different ranks and between the different types of labels of those categories.

Berlin et al. (1973) use the form of the terms for taxa as an important clue to their ranks. They distinguish primary lexemes which label life-form and generic rank taxa from secondary lexemes, which label specific and varietal rank taxa. Primary lexemes
are mostly single-word expressions that are “semantically unitary and linguistically dis-
tinct” (p. 217). They are either unanalyzable (e.g., *pine, oak, tree*) or analyzable. If ana-
lyzable, they are either productive (in which case, one of the constituents refers to the
including class, but not all members of the contrast set share the superordinate label,
e.g., *plane tree, pipe vine, lead tree*) or unproductive (in which case, no constituent
marks the superordinate class, e.g., *beggar-tick, cat-tail, poison oak, jack-in-the-pulpit*).
In contrast, secondary lexemes usually have the form “attributive + head,” in which
case one of the constituents refers to the including class and all the members of contrast
set share same superordinate constituent (e.g., *white pine, jack oak, swamp beggar-
tick*). The tricky distinction is between productive analyzable primary lexemes and sec-
ondary lexemes. Both have the form “attributive + head,” where the head term refers
to the including class. One must review the other members of the contrast set to see if
the inclusion of the head term is obligatory for all the contrasting terms.

The payoff of these fine distinctions is that one could tell the rank of a category from
its label, something that was critical in order for Berlin et al.’s (1973) theory of ranks
to work. Their nomenclatural principles worked as follows. If a taxon is labeled by a
primary lexeme and either is terminal or includes taxa marked by secondary lexemes,
it is a generic. If a taxon is labeled by a primary lexeme and includes taxa marked by
primary lexemes, it is a life-form. If a taxon is labeled by a secondary lexeme, is
immediately included in a taxon labeled by a primary lexeme, and either is terminal or
includes taxa marked by secondary lexemes, it is a folk specific. Finally, if a taxon is
labeled by a secondary lexeme, is terminal, and is immediately included in a taxon
labeled by a secondary lexeme, it is a varietal.

This was an important advance over their earlier work, for it allowed Berlin et al.
(1973) to recognize structural similarities among a large number of folk systems of bio-
logical classification. Unfortunately, there were also a number of problems with the
approach. First, there was a problem with the psychological status of the taxonomic rep-
resentation of folk biological knowledge in the first place. If your informant told you
that a *white pine* was a kind of a *pine*, that a *pine* was a kind of an *evergreen* and that
*evergreens* were *trees*, did that mean that they really had the whole structure in their
head? Randall’s (1976) answer was “No!” He argued that one could elicit a very simi-
lar chain from an informant concerning what animals eat which other animals (e.g.,
chickens eat worms, humans eat chickens, worms eat humans), but one could not con-
clude after the interview that the informant had a concept of trophic levels. Indeed,
sometimes the inclusion relations are not transitive – a *scrub oak* is a kind of *oak*, and
an *oak* is a kind of *tree*, but a *scrub oak* is a *shrub*, not a *tree*.

A second problem is that the nomenclatural principles have a number of exceptions.
For example, the term *dog* is an unanalyzable primary lexeme that labels a category that
includes a number of other categories labeled by unanalyzable primary lexemes (*hound, 
setter, terrier*), which in turn include categories labeled by secondary lexemes (*basset 
hound, Irish setter, Jack Russell terrier*). According to the nomenclatural principles, *dog* 

\[13\] The unique beginner and intermediate categories, if labeled, would violate this nomenclatural principle.
should label a life-form rather than a folk generic. The examples can be multiplied—cat, horse, cattle—all have this property. Conversely, the Awajún have a term for epiphytes, kuish ("ear"), which refers to a biological range greater than that referred to by some life-form terms (e.g., shingki "palm"), but it is not further subdivided. Dougherty (1978) presents a cogent discussion of this and other problems in folk taxonomies. A further difficulty is Gatewood's (1983, 1984) demonstration that, at least among American college students, the names of biological kinds serve as hollow place holders without much in the way of associated knowledge; the students may know the terms maple, oak, and elm as names of kinds of trees, but very often they are unable to identify any actual examples of them.

In addition to the problems mentioned above, which have received a great deal of attention, I believe that there is an even greater problem that is less well recognized. The taxonomies elicited by Berlin et al.'s (1973) approach are generally far too shallow to provide much information about the distinctions that informants readily make among biological kinds. As stated earlier, biological evolution produces very deep binary trees that usually make many feature distinctions at each node of the branching structure. In a similarity judgment task, especially one with design features of the successive pile sort [Boster (1994)], one can elicit judgments of the relative similarities and differences among a collection of specimens that captures the informants' perceptions of this binary tree structure. (A successive pile sort elicits the complete binary tree.) However, if one simply asks which life-forms the specimens belong to (presuming the specimens belong to distinct folk genera), one can only crudely approximate that binary structure. If there were 512 categories at the generic rank in a folk biological classification system, there would be 511 higher order nodes in a complete binary tree, yet there are typically at most ten categories at the life-form rank. What happened to the other 501 superordinate categories? Here again, methods derived from psychology, especially similarity judgment tasks, appear to have greater resolving power than the linguistic methods used by early cognitive anthropologists.

A final problem is the fact that Berlin et al.'s (1973) method assumes that informants agree about classification, but this assumption is patently false—informants often disagree. How do you show that folk and scientists agree on classification without assuming absolute agreement among the folk? Boster, Berlin and O'Neill (1986) provided one solution to this problem by making use of the disagreement itself to show that the folk "confuse" specimens that scientists classify as similar. In this research, bird specimens were placed on long tables and Awajún and Wambis (Aguaruna and Huambisa Jívaro) informants were asked to identify each specimen. There were two groups of bird specimens: passerines, or songbirds, comprising the largest and most recently radiated order of birds (e.g., flycatchers, thrushes, tanagers, and related species) and non-passerine birds, comprising all other orders of birds (e.g., hawks, toucans, and woodpeckers). The naming responses were used to compute an information-theoretic measure of the degree to which pairs of birds were "confused," ranging from 1 if the two bird specimens were given the same name to 0 if there was no overlap in the inventory of names applied to the pair of specimens. This measure of naming overlap was compared with taxonomic

---

14 Scientists have more explicit conventions to enforce agreement, although they are not always successful.
distance, a measure of the similarity of the birds in the scientific classification. The authors found that the Awajún and Wambis “confuse” scientifically closely related birds. Furthermore, they found that folk and scientists agree more strongly about non-passerines than about passerines. They concluded that it appears that all three groups use similar classificatory procedures to understand the natural order, but that the clarity of the natural order varies depending on the evolutionary history of the organisms. Biological classification systems are similar only to the extent that the natural order is clear.

Boster (1987) followed up this line of work and completed the shift from linguistic to psychological methods in investigating folk understanding of the biological world. The transition was necessary because the informants were chosen for their biological ignorance and, hence, did not have an elaborate set of named categories for the organisms. The article addressed the following question. It is conceivable that classifiers agree only when they have ample opportunity to study the organisms and have represented their understanding in an explicit nomenclatural system. Will subjects who have no previous knowledge of the birds sort them similarly to the Jívaro and the scientists? In this research, undergraduates with no formal training in zoology nor any familiarity with South American birds were asked to do a successive pile sort of subsets of 15 passerine specimens and 15 non-passerine specimens that had been used as stimuli in Boster (1986). Specimens were chosen such that the passerine and non-passerine subsets had the same underlying scientific taxonomic structure. Boster (1987) found that all groups of subjects (American undergraduates, Awajún, Wambis, and scientists) agreed to a considerable extent in their recognition of patterns of resemblance among the bird specimens and that, as in the earlier research, the correspondence among groups of subjects was stronger for non-passerines than for passerines. Boster (1987) concluded that these diverse human groups appear to perceive the natural order in similar ways and that recognition of that order does not seem to depend on formal training in taxonomy, intimate knowledge of the organisms, or possession of named categories for the specimens.

The next article in this series [Boster and D’Andrade (1989)] was written in response to a question posed by D’Andrade. What if the diverse groups of informants studied in the earlier research were attending to radically different features of the birds, but because the features themselves are intercorrelated, the same structure is discerned no matter what features are chosen? Agreement in classification would not tell us much about the nature of cognition if the informants could not reasonably come up with any other structure. Only if the diverse groups were choosing the same salient features of the specimens to discriminate them, would agreement on classification indicate agreement in the discernment of the organisms.

For example, imagine that the similarities of a gull, a jay, and a cardinal are judged by three informants and that the first informant judges the gull as most different because its webbed feet are different from the perching-bird feet of the cardinal and jay; the second informant judges the gull as most different because its long fish-eating beak is different from the shorter beaks of the cardinal and jay; and the third informant judges the gull as most different because its smooth head is different from the crested heads of the cardinal and jay. All three have judged the gull to be the most different, but each has
used a different basis for their judgment. It is because the salient features are correlated that all have come to the same judgment that the two passerines are more similar to each other than either is to the non-passerine.

To address this question, Boster and D’Andrade performed what they called “reverse numerical taxonomy.” Numerical taxonomists attempt to derive the best classification of a collection of specimens from a description of each specimen according to a set of features. Boster and D’Andrade reversed this process, using a set of 16 size and color measurements of two subsets of the bird specimens, to discover the relative weighting of the features in the various categorizations of the bird specimens by Awajún, Wambis, undergraduates, and scientists. Boster and D’Andrade found that the diverse groups chose the same features to differentiate each subset of specimens and chose a different set of features to differentiate the non-passerine specimens than they used to differentiate the passerine specimens. The result was interpreted as indicating that members of the different groups use similar perceptual strategies to select those features which will yield the most informative classification of the birds. Boster (1996, p. 283) argued that the human ability to categorize natural kinds demonstrated in these experiments is as much a product of natural selection as the morphological features of the organisms themselves. Summarizing this line of work, he states:

As the biological world has radiated, the capacity to recognize the order in that radiation has coevolved. The evolution of human cognition to understand the natural world is part of a more general process among living things: mind has evolved to understand nature and leaves its mark as it does so.

7. Towards a science of the stimulus

As a discipline, cognitive anthropology makes its way with a tiny fraction of the personnel of cognitive psychology. But, like other rare endemic species, cognitive anthropologists have their niche. One area of possible contribution is toward a “science of the stimulus,” which is often missing from work in cognitive psychology. The reason for its absence is fairly straightforward. Because cognitive psychology is primarily an experimental discipline, it tends to treat any properties of the stimuli as just another variable to control and manipulate and not as something that could be a possible source of explanation in its own right. (The psychologists influenced by Gibson are an important exception to this generalization.)

15 Inattentiveness to questions about the naturalness of stimuli has a long history in cognitive psychology. In Bruner, Goodnow and Austin’s (1956) classic work *A study of thinking*, many of their findings were probably the consequence of using an absolutely artificial stimulus array as their instrument for probing concept formation. The sensitivity of concept formation to the order of presentation of instances that they found was probably a result of the fact that their stimuli had no correlational structure at all. Each of the features of the stimulus array (color, number of instances, shape of figure) varied independently. The same paradigmatic contrasts of features that make for excellent experimental control also make for poor ecological validity. As discussed above, few terms in natural language label combinations of independently varying feature values unless driven by the need for constant reference to the combinations, as is the case with pronominal systems.
In contrast, because cognitive anthropologists ply their trade in natural settings, the properties of the objects of experience have played a more important role in cognitive anthropological explanations of why it is that people understand the world in the way they do. Even those cognitive psychologists who have thought hard and well about the effects that properties of stimuli presented to subjects have on cognitive performance have often not taken their ideas far enough.

For example, Rosch and Mervis (1975) demonstrated that the items in a category that are judged as most typical are those that share the most attributes with other members of the category and the fewest attributes with members of other categories. This is an important insight. However, they did not explain what process generates this particular distribution of attributes. Boster (1988) shows that the typicality ratings of birds elicited by Rosch (1975) from Berkeley undergraduates are more strongly correlated with the number of related species than with the frequency of the birds in the observers’ immediate environment or with the frequency of mention of the birds in written materials. Passerines are judged as far more typical than are members of smaller orders of birds. In other words, the pattern in undergraduates’ typicality judgments is better explained by data about the phylogeny of the birds being judged than by measures of the subjects’ own experience. Correlational structure does not emerge from thin air, but is generated by real world processes – in this case, the process of evolution by means of natural selection.

Although cognitive anthropology emerged as a discipline with a passionate commitment to the possibility of extreme cultural relativism, as the discipline has matured it has documented a number of cultural universals as well as complex patterns of cultural similarities and differences. The methods employed by cognitive anthropologists have also changed over time, as has their understanding of most of the questions addressed. The debates and dead-ends, as well as the cumulative building on each other’s accomplishments, show all the hallmarks of a healthy scientific enterprise. In general, the greatest headway has been made when the logic of the methods of data collection and analysis best match the nature of the domain studied.

The domains I have chosen to focus on in this chapter all show either some degree of cross-cultural agreement or a strong pattern in the cross-cultural disagreement. The source of agreement or the pattern in disagreement in each case appears to derive from commonalities in human experience. In the domain of kinship, those commonalities are rooted in the essential facts of life – humans are born, they mate and have children, and they die. Looking back through the generations of our ancestors are pairs of a man and a woman, each of them the son or daughter of other pairs of a man and woman, stretching backwards through time. That genealogical structure, common to all humans, is the basis of cross-cultural universals in kinship terminologies. The regular structure of the genealogical web allows certain relationships to be recognized (e.g., *uncle*, *cîtèèdèè*, *dii?*), just as the regular placement of squares on a chessboard permits the definition of possible moves of the pieces (e.g., “two over and one across” for a knight). However, there is also incredible cultural variation in how kinship is reckoned, so that the biology of reproduction appears to constrain possible kin terminologies no more severely than the use of a particular musical scale constrains the kinds of melodies that can be composed.
Agreement in color classification is more mysterious. Many of us once thought there was a neat neurophysiological explanation of the universals, but it now appears to have been illusory. My hope is that an explanation, when and if it is found, will show that the universals in color classification originate in the interaction between the characteristics of the natural phenomenon itself (the way that the spectrum of light from the sun, filtered through the atmosphere, reflects off of surfaces) and the characteristics of a human visual system that has evolved to make the best use of the information in that radiant signal. I have that hope because that is how it seems to work for human perception and categorization of biological organisms. Biological evolution has produced a marvelous pattern of similarities and differences among species and has also crafted our minds to readily make sense of that pattern.

References

Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley, CA).
Chapter 5

CATEGORIZATION IN NEUROSCIENCE: BRAIN RESPONSE TO OBJECTS AND EVENTS

CATHERINE HANSON AND STEPHEN JOSÉ HANSON

Rutgers University

Contents

Abstract 119
1. Introduction 120
2. Representing object categories in the brain 121
  2.1. Category-specific representation 122
  2.2. Feature-specific representation 123
  2.3. Process-specific representation 125
  2.4. Summary 125
3. Acquiring category knowledge 126
  3.1. Summary 130
4. Categorizing actions and events 130
  4.1. The nature of event knowledge 130
  4.2. When categorization of action fails 132
  4.3. The perception of events 133
  4.4. Summary 136
5. Conclusion 136
References 136

Abstract

Neuroscientific methods have become an increasingly important influence on the study of cognitive processing. In this chapter, we look at how the study of patient populations in addition to neuroimaging techniques have been used to address basic questions about category knowledge. How does the brain represent category knowledge? What information is acquired during category learning? Why do people parse action streams into discrete events? We examine how neuroscience has shaped the way we ask and answer questions about category learning and representation. There may not be agreement about the answers, but neuroscientific methods have helped to make investigating the questions more interesting.
1. Introduction

Whether we are recognizing a friend or making a cup of coffee, the seemingly effortless and instantaneous ability to transform sensory information into meaningful concepts determines the success with which we interact with the world. However, the seeming ease with which we engage common concepts belies the complexity of the underlying brain processing that makes categorization possible. At what point in the processing of sensory information is the category decision made? Work with nonhuman subjects suggests that categorization may be traced to individual neurons. In their seminal work, Hubel and Wiesel (1962) implanted electrodes into the visual cortex of cats and found that some neurons would respond to certain types of visual input but not to others. That is, like cells in the retina of the eye, neurons in the visual cortex of the cat appeared to be dedicated to specific types of visual input. Unlike the retinal cells, the neurons of the visual cortex responded to more complex patterns involving shape and orientation. Based on their observations, Hubel and Wiesel argued that neurons in the visual cortex were hierarchically organized into simple and complex cells, and the complex cells were constructed from more elementary ones.

The view of visual cortex as hierarchically organized was formalized by Barlow (1972) in a paper outlining the tenets of a neuron doctrine of perception. This model of perception at the neural level has dominated neuroscience studies of perception to this day. The underlying premise of the neuron doctrine is that neurons at higher brain areas are assumed to become more selective, so that “the overall direction or aim of information processing in higher sensory centres is to represent the input as completely as possible by activity in as few neurons as possible” (p. 382).

Thus, as information is propagated through the system, neurons are assumed to respond to progressively more complex and invariant features. Accordingly, it should be possible to identify neurons specialized for particular objects, for example, faces. One of the first studies to demonstrate face-responsive cells was conducted with Macaque monkeys [Gross, Rocha–Miranda and Bender (1972)]. Recording activity of single neurons in response to visual stimulation, Gross et al. found individual cells that responded maximally to face stimuli and even detected a neuron that responded maximally to a monkey hand. These specialized cells were found in the inferior temporal cortex of the monkey, which is thought to correspond, in humans, to the medial temporal lobe (MTL). Later in this chapter, we discuss data from human research that suggests the presence of face-specific brain response and its implications for category processing.

One major criticism of the neuron doctrine is that contrary to the theory, large parts of the brain can be damaged without correspondingly large changes in behavior [Lashley (1929)]. This property of the brain, which is known as cerebral mass action, and a second property, equipotentiality (the ability of a brain area to take on the function of a damaged area), suggest that brain function is not localized. Rather, these properties suggest that brain function is distributed across neuronal populations that are clustered into networks and circuits and are recruited opportunistically as various perceptual and cognitive tasks arise.
This debate between proponents of localized and distributed processing has become particularly salient in the neuroscience literature related to categorization. As we shall see, support for both sides has been collected, although recent efforts seem to favor a distributed representation of category information over ones positing localized representation. Our goal in this chapter, however, is not to argue for one view over the other, but rather to present representative research that was performed using neuroscientific techniques to study categorization.

One neuroscientific approach that has been used successfully relies on the study of patient populations. Researchers have long been interested in neurological impairments because of what they could reveal about normal cognition. Knowing that certain lesions in specific brain areas lead to fairly predictable cognitive deficits provides the means of studying the link between brain function and cognitive function. If a lesion in a given brain area produces a language deficit, it is highly likely that that area is important to normal language processing. Moreover, if that lesion does not affect other cognitive processes, it reveals something about the independence of cognitive functions.

Over the past 10 to 15 years, interest in brain function and its relation to cognitive function has led to a dramatic increase in the use of tools and techniques such as event-related potential (ERP), positron-emission tomography (PET), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) that had long been the exclusive domain of medical practitioners. By adopting the tools of neuroscience, cognitive psychologists are now able to directly (or as directly as feasible) observe the brain as it works on various cognitive tasks. Although the methodologies associated with neuroscience are not without pitfalls [see Gabrieli (1998)], cognitive neuroscientists have come very close to looking inside the black box.

The focus of this chapter is on the work that has applied the techniques of cognitive neuroscience to the study of category processing. Specifically, we examine what studies using a neuroscience approach have contributed to the understanding of how categories are learned and represented. Our review is not meant to be exhaustive, but rather to provide some examples of how neuroscientific approaches have contributed to our understanding about the processing of category knowledge in the brain.

Our discussion focuses on two types of knowledge about the world; namely, objects and actions. Most work on categorization concerns how object categories are represented and learned, and we devote the first part of this chapter to this research. A large component of human experience, however, involves the interpretation and representation of action, so we next look at how the brain processes real-world events such as *going to the theater* or *having dinner*. Finally, we offer some suggestions about future investigations.

## 2. Representing object categories in the brain

The cognitive neuroscientific approaches to category representation can be divided roughly into two camps. In one camp are those researchers who believe that there exist category-specific structures in the brain [e.g., Kanwisher, McDermott and Chun...
(1997)]. In the other camp are those who believe that category knowledge is based on distributed feature topologies [e.g., Haxby et al. (2000)]. A different approach is taken by Gauthier (2000), who sees brain structures as being linked to how category information is processed rather than to any specific category content. We review what each of these positions has contributed to our understanding of categorization.

2.1. Category-specific representation

One way in which the brain may represent categories is to dedicate a set of neurons to the task. That is, the brain could allocate particular areas to individual categories, so that whenever a category exemplar was encountered, the area of the brain dedicated to this would be activated. Moreover, brain areas dedicated to a particular category would be unresponsive when exemplars of a different category were encountered. Thus, the brain area that responded to category A would not respond to category B, and vice versa. Moreover, if category information was selectively represented in particular brain structures, damage to those structures should disrupt knowledge associated with those categories.

Category-specific disorder, a disorder in which patients demonstrate a selective dysfunction in retrieving category exemplars or features, has been associated with brain damage resulting from a wide range of etiologies including herpes simplex encephalitis, brain abscess, anoxia, stroke, head injury, or dementia of Alzheimer type (DAT). The most common form of category-specific disorder involves animate objects, inanimate biological objects, and artifacts [Capitani et al. (2003)].

Category-specific disorder often affects one type of category while sparing another. Much of the extant work in this area has found that a patient suffering from category-specific disorder will have difficulty with either the category of animate objects or that of artifacts. Rarely does a patient have difficulty with both categories. Research with patients has found that damage to areas in the inferotemporal cortex, the mesial temporolimbic structures, and the temporal pole [Gainotti (2000)] disrupts access to knowledge about living things. In contrast, knowledge about artifacts seems to be most often associated with damage in the frontoparietal areas [Gainotti et al. (1995), Saffran and Schwartz (1994)] and in the posterior left middle temporal gyrus [Martin and Chao (2001), Phillips et al. (2002)]. Figure 1 shows a picture of a brain with these areas of interest labeled.

Functional neuroimaging work looking at normal brain activation has revealed a similar pattern of results. In patient studies using PET, researchers have found that knowledge about artifacts is associated with the left posterior middle temporal area [Damasio et al. (1996), Martin et al. (1996), Moore and Price (1999), Mummery et al. (1996, 1998), Perani et al. (1995)] and knowledge about animals is associated with activation in visual association areas [Martin et al. (1996), Perani et al. (1999)]. Similar results have been found by researchers using single-photon emission computed tomography (SPECT) [Cardebat et al. (1996)], ERP [Rossion et al. (2003)], and MEG [Tarkiainen, Cornelissen and Salmelin (2002)].

Some researchers have taken this view further, arguing that categories as specific as faces [Kanwisher, McDermott and Chun (1997), Puce et al. (1995)] or buildings
Aguirre, Zarahn and D’Esposito (1998) can be shown to activate well-defined areas of the brain. Evidence for activation of category-specific brain areas rests largely on correlations between neural response and specific stimuli. For example, the fusiform face area (FFA) is so designated because it tends to respond maximally when faces are presented. Figure 2 shows the location of the fusiform gyrus.

Similarly, the parahippocampal place area (PPA) responds strongly to places. Figure 3 shows the location of the parahippocampal gyrus in the brain.

However, several researchers have questioned the existence of category-specific areas [Gauthier et al. (1999, 2000), Ishai et al. (1999, 2000)]. They offer alternative explanations of brain activation in response to category exemplars based on distributed feature networks or processing demands. We next present work that supports the existence of distributed feature networks.

2.2. Feature-specific representation

There is a growing body of work that suggests that categories may be represented by different feature-based neural systems in the brain, rather than by category-specific structures. In a review of the literature, Martin and Chao (2001, p.196) suggest that:

ventral occipito-temporal cortex may be best viewed not as a mosaic of discrete category-specific areas, but rather as a lumpy feature-space, representing stored information about features of object form shared by members of a category.
This “lumpy feature-space” can be thought of as a feature-map [Gauthier (2000)] or object-form topology [Haxby et al. (2000)].

Much of the evidence against category-specific brain structures comes from work showing that the response of these areas is not restricted to specific categories alone [e.g., Blonder et al. (2004), Haxby et al. (2001)]. Chao, Haxby and Martin (1999) found that the lateral fusiform gyrus and the right posterior superior temporal sulcus responded to animals, faceless animals, and human faces.

Recent findings in our lab [Hanson, Matsuka and Haxby (2004)], in which we used a neural net classifier to model brain activation patterns, do not support the notion of category-specific brain structures. We used a voxel-wise sensitivity analysis to look at category related responses in the ventral temporal (VT) lobe. Our results indicate that the same VT lobe voxels contributed to the classification of multiple categories, which does not support a category-specific model. Instead, we argue for a combinatorial coding of category features in VT.

The advantage of a distributed representation of features approach is that it can account for activation of the same brain structure when exemplars from different categories are present. On the other hand, this approach is not good at explaining why a
brain structure that responds when certain objects are present does not similarly respond when other objects that have similar features are present [Gauthier (2000)]. Objects that share similar features should be expected to elicit similar activation in the same brain structure, but this is not always the case [Epstein and Kanwisher (1998)].

A third account of category representation in the brain, the process-map model [(Gauthier (2000)], takes an entirely different approach. In this model, neither categories nor features per se are represented, but instead brain structures are seen as computationally specialized for category related processing. We look at this approach next.

2.3. Process-specific representation

Gauthier (2000) argues that brain response to category information is based not on category content, but rather on category process. In her words (2000, pp. 1–2), “extrastriate cortex contains areas that are best suited for different computations.” Category exemplars elicit brain activation based on how the information is to be used (e.g., level of categorization) as well as on prior experience. Consequently, the process map reflects experience and processing goals.

In a direct challenge to the category-specific approach, Gauthier (2000) argues that brain response to faces may seem to differ from that for objects, because faces are processed at a more specific level of categorization than are most objects. Moreover, she argues, within-category discrimination may be more relevant for faces than for other objects, leading to a greater expertise in selecting critical features.

To examine the effect of expertise on FFA, Gauthier et al. (1999) trained study subjects in a category comprising of artificially generated exemplars, which she called “greebles.” She used fMRI to scan subjects first before they had any experience with the greebles, then at three different times during the learning phase, and finally, twice after they reached a learning criterion. Gauthier then examined the region of interest (ROI) previously associated with face processing, the middle and anterior fusiform gyri. She found the expected response in the presence of faces. She also found, however, changes in brain response as the study participants’ experience with greebles increased. Gauthier concludes that the fusiform gyrus, rather than being a category-specific structure of the brain associated with faces, is the site of fine-level category discrimination that changes as a function of experience.

Evidence for the fusiform gyrus being activated during fine-level category discrimination was also found by Tyler et al. (2004).

2.4. Summary

The studies of category representation that we have reviewed in this section provide three potential accounts of how category knowledge is represented in the brain. Although they are very different, they are not mutually exclusive. It is certainly feasible that areas of the brain might respond differentially to different objects, even if category knowledge is represented by a distributed feature topology [Martin and Chao...
Moreover, any model of category representation must account for learning and within-category discrimination, and it is quite possible that the distributed feature topology underlying category representation varies as a function of the processing demands needed for a given task [Tyler et al. (2004)]. There is also the possibility that category representations may differ across the two brain hemispheres, with the left hemisphere representation being more category-specific and the right hemisphere representation being more similar to a feature topology [Deacon et al. (2004)].

While a definitive account of category representation remains to be determined, the direction that research needs to take about the neural substrate of category representation is becoming clearer. In the next section we look at studies that focus on how categories are learned. The questions are different but the goal of understanding the role of the brain in category knowledge is similar.

3. Acquiring category knowledge

Objects and entities may be grouped together, i.e., form a category, for many reasons. A category such as red objects is a relatively simple category whose members include red things. The red things may be small or large, animate or inanimate, soft or hard, noisy or quiet, etc. It is only necessary that the members share the property of redness. A different kind of category, for example the category game, cannot be easily defined by a set of necessary and sufficient features. A game may be played by one or by many, it may or may not be played outdoors, and scoring may or may not be required, etc.

Categories such as red objects are known as well-defined categories, because they are completely defined by necessary and sufficient features. On the other hand, categories such as game are known as ill-defined categories, because they are not readily defined by necessary and/or sufficient features. Thus, learning a rule that includes the features necessary for category membership may be sufficient to correctly categorize members and nonmembers of a well-defined category. If a concept is ill-defined and does not have necessary and sufficient features, however, the basis of its representation is not as clear cut.

Categories that cannot be learned on the basis of a rule may be learned through similarity judgments between the to-be-categorized object and exemplars that have already been identified. This can be accomplished in two ways. First, a similarity judgment can be made by comparing one or more exemplars with the to-be-categorized object. Second, a similarity judgment can be made by comparing some central tendency or prototype generated from experience with exemplars to the to-be-categorized object.

Earlier, we examined how category information might be represented in the brain. In this section, we look at how the brain responds when categories are being learned. Concept learning involves learning what it is that binds exemplars of a category together. Our consideration of brain response during concept learning will focus on a traditional distinction made by cognitive psychologists between explicit, analytic processing and implicit, nonanalytic processing.
Explicit, analytic processing is associated with learning based on rules that specify the necessary and sufficient features required for category membership [Bruner, Goodnow and Austin (1956)]. Analytic processing occurs most often with well-defined stimuli, in which category membership depends on specific features shared by all members of the category. Exemplars of categories that are rule-based do not differ in the degree to which they are typical of a category inasmuch as they differ in the degree to which all members must share the necessary features.

Implicit, nonanalytic processing involves similarity-based comparisons between the to-be-categorized object and a representation of the candidate category. The representation that is used in the similarity judgment may be either known members of the candidate category or a derived prototype of the category. Similarity judgments based on exemplars rely on the similarity between a to-be-categorized object and the exemplars of a candidate category. Thus, one may learn about dogs by noting the similarity between a newly encountered dog and a particular dog such as the poodle next door, or the family basset hound.

Alternatively, similarity judgments during category learning may rely on a prototype. A prototype is generated through experience with individual exemplars, yet is different from any given exemplar. The generation of a prototype occurs without conscious effort and often, is not easily described in words. Category structure based on a prototype is most similar to family resemblance [Rosch and Mervis (1975)], in which some members are more typical than others.

Do analytic and nonanalytic processing reflect different neural structures? Work with individuals with brain damage has provided some indication that different neural structures underlie different types of category processing. Damage to the medial temporal lobe (MTL) structures that include hippocampus and amygdala is associated with anterograde amnesia, a dysfunction characterized by the inability to explicitly access newly acquired information. Individuals with anterograde amnesia are capable of learning new information, but lack the ability to consciously or explicitly access it. Consequently, medial temporal (MT) brain structures have been associated with declarative memory processes, those responsible for processing factual, explicitly available information.

In contrast to declarative knowledge, procedural knowledge such as that involved in skill learning does not require explicit or conscious retrieval. Although patients with anteretrograde amnesia cannot access information for tasks that require explicit processing, they can learn new sensorimotor skills such as mirror tracing [Milner (1962)] or rotary pursuit [Corkin (1968)]. Patients who are most impaired in sensorimotor skill learning are those with diseases that affect basal ganglia, especially Huntington’s disease (HD) and Parkinson’s disease (PD).

This difference between analytic and nonanalytic processing is also found when brain-impaired individuals engage in probabilistic classification tasks. Patients with amnesia tend to show normal learning during the early but not later learning stages of classification problems [Knowlton, Squire and Gluck (1994)]. However, HD and PD
patients tend to be impaired throughout this learning process [Knowlton, Mangels and
Squire (1996)]. Thus, it appears that MTL and basal ganglia support different aspects
of processing during probabilistic classification.

Further support of the localization of two kinds of category processing was obtained
by Myers et al. (2003). In this study, patients with MT amnesia and patients with PD were
first exposed to pairings in which two stimuli A1 and A2 were both associated with a third
stimulus X1. They were then shown pairings of A1 with a new stimulus X2. Myers et al.
aimed to determine whether or not presentation of the stimulus A2 would lead the study
participants to pair that stimulus with X2, thereby demonstrating that they had equated
A1 and A2. The investigators predicted that they would observe a double dissociation in
learning performance between study participants suffering from hippocampal atrophy
and those suffering from PD. This was confirmed. They conclude that basal ganglia are
responsible for simple associative learning early on and the hippocampus is responsible
for the ability to transfer knowledge to new exemplars. The finding that patients with
amnesia have difficulty with more complex associations is consistent with the results of
an earlier study by Kolodny (1994) who found that individuals with amnesia were able
to learn simple, perceptually based categories, but had difficulty with more complex stim-
uli. Neuroimaging studies with healthy individuals support this distinction between the
structures that underlie analytic and nonanalytic processes during a categorization task.
Smith, Patalano and Jonides (1998) scanned experimental participants with PET while
they performed either an exemplar- or a rule-based categorization task. They found
frontal activation of the dorsolateral prefrontal cortex (DLPFC) and areas of the frontal
cortex to be uniquely associated with the rule-based task while the exemplar-based task
uniquely activated the left visual cortex (Brodmann area 18) and the left cerebellum.

Support for distinct brain structures underlying analytic and nonanalytic processing has
also been found in fMRI studies. Poldrack et al. (2001) found activation of MTL to be neg-
atively correlated with activity in basal ganglia during a category learning task. They argue
that MTL may be most involved whenever flexible accessibility to knowledge is needed,
whereas striatum (a subset of basal ganglia that includes the caudate, the putamen, and the
nucleus accumbens) supports fast, automatic responses. Drawing on both animal and
human research, Poldrack and Packard (2003) conclude that MTL and basal ganglia func-
tion as independent memory systems that are capable of interacting with one another.

Lieberman et al. (2004) also found a dissociation between MTL and caudate in an
fMRI study of artificial grammar learning. In the Lieberman et al. study, MTL was
associated with “chunk strength,” a measure of the similarity between test and training
items, and the caudate was associated with “rule adherence,” or grammaticality. By
manipulating both the chunk strength and rule adherence of items independently,
Lieberman et al. were able to assess the contribution of the MTL and the caudate to
learning, and to determine the interaction between nonanalytic (chunk strength) and
analytic (rule adherence) processes. They found that the MTL and the caudate were
negatively correlated, as was reported in the Poldrack et al. (2001) study.

Similar dissociations using the artificial grammar learning paradigm have been
found between the MTL and the prefrontal cortex in healthy adults [Opitz and
Friederici (2003, 2004)] as well as in patient populations [Ullman et al. (1997)]. Taken together, these studies suggest that grammar learning progresses from similarity-based processing in the MTL to more complex processing in the prefrontal cortex.

A direct investigation of analytic and nonanalytic processes and the associated neural structures involved was performed by Tracy et al. (2003). They instructed participants to classify pseudowords on the basis of a criterion attribute (CA) or on the basis of family resemblance (FR). CA processing was found to be associated with the right anterior temporal and the inferior frontal regions of the brain, whereas FR processing activated the medial cerebellar and the left extrastriate areas.

In our laboratory, we found that study participants who are biased toward a configural orientation (making a judgement about expressed emotion) will perform differently on integral (condensation) and separable (filtration) categorization tasks than will those who are biased toward a featural orientation (making a judgement about facial features) [Hanson, Hanson and Schweighardt (2004)]. Specifically, we found that feature level processing significantly impeded a subsequent categorization task based on integral features but not one based on separable features (see figure 4).

We imaged experimental participants with fMRI as they performed the category task and found parahippocampal gyrus and posterior cingulate to be active during categorization when the task was inconsistent with the initial orientation (see Figures 5a and 5b, Figure 5a and 5b is Plate 5.5a and 5.5b in the Separate Color Plate section).

![Fig. 4. Left-hand barplot shows the correct reaction time (RT) from overall subjects. Note modulation of condensation by priming condition.](image-url)
3.1. Summary

The evidence from both patient studies and neuroimaging studies of healthy individuals suggest that analytic and nonanalytic category learning engage different neural structures. Similarity-based categorization is very likely accomplished in the MTL, whereas rule-based categorization apparently engages the basal ganglia. Complex rule learning, such as that involved in learning artificial grammars, may also involve the prefrontal cortex. The prefrontal cortex may also play a major role in the categorization of action units or events. Although much of the cognitive neuroscience work on categorization has focused on object categorization, interest in how the brain parses action streams into discrete and meaningful units is increasing. We next look at some of the research in this field.

4. Categorizing actions and events

Our discussion of categories thus far has focused on object categories. Yet, a very important type of category knowledge concerns the way individuals categorize the dynamic information associated with actions and activities. We recognize someone driving a car, we arrange to have lunch with a friend, and we relate our hopes about moving to a new house. Whereas object categories lend themselves to representations based on perceptual or functional qualities, the perceptual attributes associated with events such as making a phone call are not as well defined. The person making the call, the type of telephone being used, the time of the day, the location, the duration of the call, and the target of the call are not constrained in the same way that the attributes of apple or table are.

4.1. The nature of event knowledge

When people remember vacationing in Provence and tell each other about what happened at work, they are imposing a categorical structure on what is essentially unstructured,
continuous activity. Without the shorthand afforded by event categories, it would be an impossibly tedious task to recount the actions involved in a single morning; the time needed to recount performed actions would approach the duration of the actions themselves. Even the seemingly endless recounting of some favorite memory by a too verbose friend does not approach the duration of the actual experience of the events.

It is this element of time that distinguishes event categories from object categories. Whereas objects occupy space, events occupy time. Recognition of objects is immediate under most circumstances, but recognition of events must evolve in time. Perhaps most tellingly, temporal order can change the meaning of events, whereas spatial order rarely affects the meaning of objects.

Despite these differences, event categories share many basic features of category representation with object categories. Like many object categories, event categories have a typicality structure. Reading a menu is a fairly typical exemplar of the class of actions involved in going to a restaurant, whereas a waiter singing to the customers is not. It is this typicality structure that allows the listener to infer the events being seated, reading a menu, ordering food, paying a bill when the speaker relates that she went out to dinner.

Similarly, event categories, like object categories, have a partonomical structure [Miller and Johnson-Laird (1976), Tversky (1990), Tversky and Hemenway (1984), Zacks and Tversky (2001)]. That is to say, an event is composed of other events. In much the same way that a table is made up of a surface and legs, the event setting the table involves carrying plates and utensils to the table and arranging them appropriately. Setting the table itself may belong to other events such as throwing a party or feeding the family. Objects, unlike events, also have a taxonomic structure, so that category members inherit the properties of the superordinate category. Partonomies do not propagate the properties of a superordinate to subordinate levels. Although buying a ticket is part of going to the theater, the features of going to the theater are not propagated to buying a ticket in the way that the features of an animal (such as breathing and eating) are propagated to members of the animal category. If a mouse is an animal, it breathes and eats. In contrast, how one attends a theater performance reveals little about how one buys a ticket.

Although there are many different ways to buy a ticket (at the theater, through the mail, from the Internet), the event itself, buying a ticket, is a highly probable component of attending the theater. Thus, expectations about event categories are primarily functional, rather than perceptual. If you are asked to retrieve an apple from the kitchen, you have certain expectations about the size, color, and shape associated with the apple category that allow you to choose an apple when you see it rather than selecting an orange or banana. Expectations about events are not bound by the same kind of perceptual constraint. So, how do we categorize dynamic action sequences into distinct events?

We have argued elsewhere [Hanson and Hanson (1996)] that the perception of events is a cyclical process in which memory focuses attention, attention selects information from the environment, and attended information influences the activation of memory. Using a recurrent neural network to simulate human event judgments, we were able to demonstrate that this kind of perceptual cycle [Neisser (1976)] is a viable account of how
people categorize events. Moreover, the expectations (memories) that guide attention include a temporal element. In other words, the duration of an event is one of its critical features and is used to segment action streams into individual events.

4.2. When categorization of action fails

Although we take for granted our ability to interpret and produce mundane events such as making coffee or mailing a letter, failure to comprehend or perform simple action sequences can have a debilitating effect on daily life. The inability to imitate or perform an action sequence on demand, despite intact sensory and motor ability, is labeled apraxia. A broad distinction can be made between conceptual (or ideational) apraxia and production (or ideomotor) apraxia [Liepmann (1920)]. That is, the source of impairment may rest in the action representation (conceptual) or in the execution of action (production).

Evidence for different neural circuits underlying conceptual and production apraxia is limited, although some indication of distinct systems has been obtained. For example, Rapcsak et al. (1995) studied a patient with a progressive bilateral limb apraxia who complained of severe spatiotemporal problems related to skilled limb movement that were confirmed by clinical assessment and a three-dimensional computerized graphic analysis of her movements. Despite severe impairment in the ability to produce action, the patient’s conceptual knowledge of action seemed relatively intact. Neuroimaging by MRI revealed atrophy of the posterior-superior parietal lobes, with relatively normal frontal, temporal, and occipital areas. Scanning with SPECT found extensive posterior cortical dysfunction involving the temporoparietal area, which on the left side extended into the inferior parietal lobule. Conceptual apraxia often accompanies damage to the left posterior parietal and/or the premotor cortex.

In a review of lesion studies and neuroimaging work, Johnson-Frey (2004) concludes that the different types of impairment seen in conceptual and performance apraxia stems from functionally specialized neural systems underlying semantic knowledge and procedural knowledge. Specifically, he provides an account involving distributed neural systems in the left temporal, parietal, and frontal areas. Based on a variety of studies, he suggests that parietal areas are involved primarily with action performance, temporal areas with action knowledge, and frontal areas with both production and representation of action.

Much of the work on apraxia focuses on a subcategory of action knowledge concerned with skilled performance and tool use. Tool use, however, is only one aspect of how people use action categories. People continually parse a stream of activity into meaningful units. The ability to parse action sequences into meaningful units underlies our ability to make sense of the world and to communicate with others.

Patients with schizophrenia are another population who demonstrate difficulty with action categorization. A primary symptom of schizophrenia is the incorrect attribution of self-generated action to external sources. This misattribution often leads to delusions in which the individual believes others are controlling his or her thoughts and/or actions. There is some evidence that this dysfunction is related to the inability of schizophrenic
individuals to differentiate intention from recognition of action [Daprati et al. (1997)] or
to represent predicted consequences of action [Frith, Blakemore and Wolpert (2000)].
The latter authors suggest that a disconnection between frontal brain areas that initiate
action and parietal areas that represent body state is responsible for the misattribution
of self-generated action. They note that in healthy individuals, brain activity in the frontal
areas is inversely proportional to that in the posterior areas, but in schizophrenic indi-
viduals activity in the frontal and posterior areas are independent. Frith et al. suggest that
this disconnection between the anterior and posterior cortex indicates that the response
to self-generated actions are not suppressed in schizophrenic individuals, unlike in
healthy individuals.

Schizophrenic patients also demonstrate difficulty in retrieving action knowledge
et al. (2001) found that schizophrenic patients, even more than patients with frontal lobe
afflictions, had trouble sequencing actions and prioritizing actions related to a goal.
Zalla et al. (2004) asked schizophrenic patients and healthy controls to parse action
sequences into small and large units. Compared to the control subjects, the schizo-
phrenic subjects had difficulty detecting large action units; this difficulty was positively
correlated with higher levels of disorganization in these subjects. Zalla et al. also found
that schizophrenic patients remembered the action information differently than did
healthy subjects. Schizophrenic patients were more likely to recall actions without
regard to temporal order, to personalize the characters in videos, and to confabulate.

Unlike schizophrenic individuals, healthy individuals rarely have difficulty distin-
guishing between self-generated and other-generated action. This is particularly notable
given the growing evidence that similar neural systems underlie observation, mental
simulation, and imitation of actions [see the review by Decety and Chaminade (2003)].
The prefrontal cortex, anterior cingulate, premotor cortex, inferior parietal lobule, ven-
trolateral thalamus, and caudate appear to be involved in the representation of self-gen-
erated action as well as action performed by others. Distinguishing self-generated
action from that generated by an external agent appears to be accomplished in the infe-
rior parietal cortex and the right prefrontal cortex [Decety and Chaminade (2003)].

4.3. The perception of events

Healthy individuals effortlessly parse action streams into discrete units (events) and
moreover, demonstrate a high degree of consensus in judgments of event boundaries
[Hanson and Hirst (1989), Newton,Engquist and Bois (1977)]. We use events to under-
stand, remember, and communicate what would otherwise be an overwhelming and
chaotic deluge of stimulation. Although considerable effort has been spent studying
how object categories are represented and learned, much less attention has been allo-
cated to understanding how the brain makes sense of the what William James calls the
“blooming, buzzing confusion” that is everyday experience.

How does the brain categorize information that extends both in time and space? This
was the question posed by Zacks et al. (2001), who used fMRI to scan the brains of
healthy subjects while they watched short videos of common activities such as making the bed and doing the dishes. Looking at brain activity during event transitions, they found activation in the bilateral posterior cortex and the right frontal cortex, with prominent activity located in the occipital/temporal junction, at the middle temporal/ventral 5 (MT/V5 complex) also known as the extrastriate motion complex. The MT/V5 area is associated with processing biological motion and human action, and the right frontal area is associated with active shifts of attention and eye movements.

A subsequent study of event perception by Speer, Swallow and Zacks (2003) used fMRI in an ROI analysis of the MT/V5 complex and the frontal eye field (FEF). These areas were chosen because the MT/V5 is known to respond to visual motion and the FEF is known to orient eyes to a visual stimulus through guided saccadic and smooth pursuit eye movements. Speer et al. reasoned that if motion changes contribute to the perception of event boundaries, then activity in these areas should correlate with judgments about event transition points. They conclude that event perception is related to the detection of visual changes, but leave open the question of whether MT/V5 and FEF are driven directly by bottom-up (stimulus-based) activation or modulated by top-down (knowledge-based) processes.

Some evidence for top-down processing during event perception [Sitnikova, Kuperberg and Holcomb (2003)] has been found using ERP. Sitnikova et al. had subjects watch short videos of common events such as shaving or cooking. The key manipulation was whether an appropriate or inappropriate object was used in the sequence (e.g., whether a razor or a rolling pin was used in the shaving sequence). They found that a strong negative ERP (N400) accompanied the appearance of the critical object, and they interpret this result as indicating that a strong temporal relation exists between object identification and scene comprehension during event perception.

Indirect evidence for top-down processing during event perception has also been found in recent work by Wood et al. (2003). They used fMRI to look at brain activity in healthy subjects as they categorized words and phrases associated with social and nonsocial structured event complexes (SECs). SECs are familiar event sequences that others have labeled as “scripts” or “schemas.” The fMRI data revealed differences in activation patterns between social and nonsocial conditions in the prefrontal cortex (PFC), which Wood et al. interpreted as evidence for category-specific representations in the PFC.

In a recent study in our laboratory [Hanson and Hanson (2005)], we manipulated bottom-up and top-down factors directly during an event perception task. We scanned the brains of study subjects with fMRI while they watched either highly familiar, common events such as drinking coffee, or a novel cartoon depicting a geometric shape moving around other geometric shapes. Figure 6 shows the temporal response density (TRD) for the real-world familiar event; the 99% confidence limit is shown by a frequency response near 15. At this threshold most subjects identified 11 independent event change points in the video sequence.

Using these TRD event points we constructed a contrast between event change points and non-event change points. Figure 7 (Figure 7 is Plate 5.7 in the Separate Color Plate section) shows the analysis of an fMRI scan of the brain of a subject parsing a
Fig. 6. Temporal response density (TRD) for subjects watching a video of a student studying. The points above the threshold (15 responses) represent significant agreement across subjects.

Fig. 7. Contrast between event change points determined by significant temporal response density (TRD) points and nonevent points. Typical areas that appear engaged are the anterior cingulate, the inferior frontal gyri, the middle frontal gyri, the precuneus, the parietal lobule, and the dorsal lateral prefrontal cortex. These are voxel activities that are used for further clustering and graph analysis.
real-world familiar sequence. Typical areas of activity that are found include the bilateral prefrontal areas, the anterior cingulate, and other related attentional areas.

We cluster the areas using a mode-density clustering algorithm that tends toward sparse cluster structure and determines \(<x,y,z>\) centroid voxels for each ROI. These centroids are then used to calculate covariance between ROIs to determine their interactivity by finding the best-fitting graph using LISEREL. Figures 8a, 8b and 8c are best-fit graphs for doing a visual oddball task (detection task with tonic background), parsing a simple geometric stimulus, and parsing a real-world familiar task, respectively. Note that the oddball task contains some constituent ROIs (the inferior frontal and the middle frontal gyri) that also appear in the event–parsing tasks. This might suggest that various areas are recruited as constituents in a larger computational network.

4.4. Summary

Studies of event perception, such as those we have reviewed here, provide a unique opportunity to observe the dynamic interaction of brain areas during the performance of a real-world cognitive task. Parsing action streams into discrete, meaningful units involves the recognition of not only object boundaries, but also action boundaries. Because events take place in time, sequence information is important. Studies of patients with apraxia and patients with schizophrenia suggest that the frontal lobe plays an important role in processing sequence information. In addition, the neuroimaging studies implicate both MT/V5 and prefrontal areas as being involved in event perception.

5. Conclusion

Categorization is a fundamental cognitive skill. Due to the use of cognitive neuroscience methods, our understanding about both the representation and processing of
category information has been greatly advanced. It is clear that category learning and representation requires interaction between a number of brain areas that are constitute processes of other complex cognitive functions like language, executive function, and attention. Future understanding of basic cognitive processes will likely engage both object and event representations to determine the nature of fundamental brain processes underlying these common complex categorization functions.

References


dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that
grammatical rules are processed by the procedural system”, Journal of Cognitive Neuroscience
127:3–21.
processes in schizophrenia: A comparison with patients with frontal lobe damage”, Cognitive
Abstract

In cognitive science, computers can be used in three ways: to simulate cognition for artificial intelligence (AI), to enhance cognition by assisting human intellectual activity, and to help scientists understand cognition by testing theories on large amounts of data. These three approaches are not mutually exclusive, since specialists in any of these areas frequently adopt techniques designed for the others. This chapter surveys theories of categorization and reasoning in cognitive science that have been implemented and tested in computer systems. Most of the ideas originated long before modern computers were invented, but computers provide an opportunity for developing them in greater detail than was previously possible. Section 1 presents early work in the field of computation in cognitive science. Section 2 surveys top-down and bottom-up approaches to categorization; Section 3 analyzes the implications of structure, context, and purpose on the choice of categories and the methods for recognizing individuals that belong to those categories; and Section 4 considers the interactions of categorization and reasoning. The concluding Section 5 discusses the levels of cognition and some successes and failures in simulating those levels by AI.
1. Computation in cognitive science

Theories of categorization in artificial intelligence (AI), information retrieval, data mining, and other computational fields are no different in kind from theories that predate modern computers. The computer, however, introduces two important elements: it enables theories to be tested on large amounts of data, and it enforces precision, since no program running on a digital computer can ever be vague or ambiguous. Both elements can be helpful in formulating and testing theories, but neither can guarantee truth, relevance, or usefulness. Sometimes, as Lord Kelvin observed, precision can be a distraction: “Better a rough answer to the right question, than an exact answer to the wrong question.” To avoid a bias toward answers that are easy to program, it is important to consider questions posed before computers were invented.

This chapter surveys a variety of computational methods that have been applied to categorization and related methods for reasoning with and about the categories. These methods can be used for three different purposes:

1. **Artificial intelligence (AI).** From the earliest days, computers were considered “giant brains,” which had the potential to mimic and perhaps surpass human intelligence. Good (1965) predicted “It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built and that it will be the last invention that man need make.” Except for chess playing, that prediction has not come to pass, but attempts to achieve it have contributed to a better appreciation of the richness and complexity of human intelligence.

2. **Intelligence enhancement.** Computer capabilities are very different from and complementary to human abilities. That difference has led to a wide range of tools that supplement human cognition: information storage, management, search, query, and retrieval; text editing, analysis, translation, formatting, and distribution; calculation in graphics, spreadsheets, statistical packages, and formula manipulation; and computer-aided human communication and collaboration.

3. **Hypothesis testing.** Psychology, linguistics, and philosophy deal with complex phenomena that cannot be described by the elegant mathematics used in physics. Without computers, theorists are often limited to studying an unrepresentative sample of oversimplified examples or to collecting statistics that show trends, but not causal connections. A computer, however, can analyze large volumes of realistic data and test hypotheses about causal mechanisms that could generate or interpret such data. These three approaches differ primarily in their goals: simulation, enhancement, or understanding of human cognition. Computational methods designed for any one of them can usually be adapted to the others.

Neural networks illustrate the way theories introduced before modern computers have contributed to computer-based techniques. In his study of learning, Thorndike (1932) developed a stimulus–response (S–R) theory, which he called connectionism: rewards strengthen the S–R connections, and punishments weaken them. Meanwhile, McCulloch and Pitts (1943), a neurophysiologist collaborating with a logician, designed a theoretical model of neural nets that are capable of learning and computing any Boolean function of...
their inputs. To implement that model, Rosenblatt (1958) built a machine called a perceptron, which simulated the nodes and links of a neural net. Later versions combined connectionism with neural nets, which were simulated on digital computers instead of being implemented directly in hardware. Figure 1 shows a typical neural net with stimuli entering the layer of nodes at the left; the links propagate signals to nodes in one or more hidden layers in the middle; finally, the responses appear at the layer of nodes on the right. The nodes represent neurons, and the links represent dendrites that pass signals from one neuron to another. In computer simulations, each link has an associated weight, which increases or decreases the strength of any signal that passes through the link. Each node combines the strength of all its inputs to determine the value of its output.

Behaviorist methods of operant conditioning were adapted to neural-net methods of learning by backpropagation. Whenever a network generates incorrect responses for given stimuli, it is, in effect, “punished” by adjusting the weights on the links, starting from the response layer and propagating the adjustments backwards toward the stimulus layer. Over the years, many variations have been proposed for functions that combine inputs to determine outputs, for methods of backpropagation, for algorithms that use and change the weights, for the number of nodes and hidden layers, and for possible feedback from response layers to earlier stimulus layers. All these variations are incremental modifications of an older theoretical foundation.

The example of neural nets is interesting in its own right, but it also shows how a theory can be used for different purposes and how the same aspects may sometimes be an advantage or a disadvantage:

- In 1943, neural nets were proposed by McCulloch and Pitts as a model of how the brain works. But as more data about neurons became available, the original theory proved to be grossly oversimplified. It was, however, a valuable first step that clarified the issues and helped subsequent researchers design more realistic models.
- Although the original version was no longer adequate as a model of the brain, it continued to be used as a valuable computational technique for many kinds of pattern recognition.

![Fig. 1. A neural net for connecting stimuli to responses.](image-url)
recognition. Instead of a model for understanding intelligence, it became a method of enhancing intelligence.

- Rosenblatt (1958) called his perceptron a “probabilistic” model, which is a more accurate description than the term neural net, since actual neurons are much more complex than the simple nodes of Figure 1. The word neural gives the model an unwarranted mystique that has misled casual observers and sometimes even people who should know better.

- The behaviorists viewed S–R mechanisms as a “black box” that avoided talk about concepts and other “mentalistic” terms. Yet that black-box quality, which the behaviorists considered an advantage, is one of the greatest weaknesses of neural nets: there is no way to explain or justify their responses. Although a programmer can look inside the nets, there is nothing to see but a meaningless jumble of numbers.

- McCulloch and Pitts (1943) treated each neuron as a simple switch, which was easy to simulate, but current evidence indicates that a neuron is more like a one-chip computer with its own local memory and the ability to receive, transform, and transmit complex signals. How such processes should be simulated is still a major research question that remains to be answered.

The remainder of this chapter examines various ideas, themes, and theories of cognitive science that have been adapted and applied in computer systems. In most cases, the ideas originated long before modern computers, but the ability to process large volumes of data with great precision has led to a depth and sophistication that was previously unattainable.

2. The great categorization debates

Aristotle invented the two dominant methods of classification, which in different variations have been used in all branches of cognitive science. In his logical works, he presented a top-down method for defining concepts based on a genus or supertype and one or more differentiae, which distinguish the new type from others of the same genus. But in his biological works, he criticized the limitations of the top-down approach and recommended a bottom-up approach beginning with a detailed description of individuals, classifying them in species, and grouping species in genera. He considered the top-down method appropriate for presenting the results of analysis and reasoning about them, but he recommended the bottom-up method as a better discovery procedure for investigating a new subject.

In the third century AD, the philosopher Porphyry wrote a commentary on Aristotle’s categories, which contained the first recorded tree diagram (Peter of Spain 1239/1947). The version in Figure 2 illustrates the categories and their relationships to syllogisms, which are Aristotle’s rules for reasoning about types and subtypes. With the differentia material, the supreme genus Substance becomes the subtype Body, and with the differentia immaterial, becomes Spirit. The technique of inheritance is the process of merging the differentiae along the path above any category: LivingThing is defined as “animate material Substance,” and Human is “rational sensitive animate material Substance.”
The goal of mechanically relating concepts to primitives was first proposed by Ramon Llull (1303) in the thirteenth century. His *Ars Magna* was a system of disks inscribed with primitive concepts, which could be combined in various ways by rotating the disks. Lull’s system inspired Leibniz (1666) to develop his *Universal Characteristic*, which represented primitive concepts by prime numbers and compound concepts by products of primes. Then statements of the form “All A is B” are verified by checking whether the number for A is divisible by the number for B. If Plant is represented by 17 and Deciduous by 29, their product 493 would represent DeciduousPlant. If Vine is represented by 20,213, the statement “All vines are deciduous” is judged to be true because 20,213 is divisible by 493. Leibniz envisioned a universal dictionary for mapping concepts to numbers and a calculus of reasoning that would automate the syllogisms. To simplify the computations, he invented the first calculating machine that could do multiplication and division.

With the advent of electronic computers, computational linguists set out to implement Leibniz’s universal dictionary. For her machine translation system, Masterman (1961), a former student of Wittgenstein’s, defined a *semantic net* of 15,000 words, organized as a lattice based on 100 primitive concepts, such as [FOLK], [STUFF], [CHANGE], [GO], and [TALK]. For the sentence “This man is greedy, but pusillanimous,” her system would generate the representation

\[
(\text{THIS: MAN:}) \ (\text{HE:} \ (\text{CAN/ DO/} \ (\text{MUCH: EAT:}))) \\
(\text{BUT: NOT:}) \ (\text{HE:} \ (\text{CAN/ DO/} \ (\text{MUCH: FIGHT:}))).
\]

Fig. 2. Tree of Porphyry, translated from a version by Peter of Spain (1239/1947).
For conceptual dependency graphs, Schank (1975) reduced the number of primitive acts to 11. The phrase “x bought y,” for example, could be expanded to “x obtained possession of y in exchange for money.” Transforming high-level concepts to primitives may show that two different phrases are synonymous. But many deductions are shorter and simpler in terms of a single concept like [LIAR] than a graph or formula for “one who mentally transfers information that is not true.” In general, a system should allow high-level concepts to be expanded in terms of lower ones, but such expansions should be optional, not obligatory.

Modern dictionaries analyze thousands of words into more primitive ones, but they are not limited to a fixed set of categories. They also allow circular definitions, such as defining attribute as characteristic and characteristic as attribute. In linguistics, Katz and Fodor (1963) introduced primitives called semantic markers with projection rules for combining them. Many linguists adopted variations of the Katz–Fodor method, but even those who used it raised serious criticisms:

- No linguistic or psychological data provides evidence for the sharp break between the information in the semantic markers and the leftover information, which Katz and Fodor called the distinguisher.
- Most languages contain families of synonyms like glad, happy, cheerful, light-hearted, and joyful, each with a slightly different shade of meaning. But semantic markers only support either/or dichotomies.
- Semantic markers, like Leibniz’s prime numbers, can only represent conjunctions of primitives. Other operators are needed to represent all logical relationships.

All logic-based methods ranging from the Tree of Porphyry to the latest formal ontologies are examples of an Aristotelian top-down approach, but Aristotle himself recommended bottom-up methods for analyzing empirical data. Whewell (1858) went further in claiming that top-down definitions are useless in biology:

Natural groups are given by Type, not by Definition. And this consideration accounts for that indefiniteness and indecision which we frequently find in the descriptions of such groups, and which must appear so strange and inconsistent to anyone who does not suppose these descriptions to assume any deeper ground of connection than an arbitrary choice of the botanist. Thus in the family of the rose tree, we are told that the ovules are very rarely erect, the stigmata usually simple. Of what use, it might be asked, can such loose accounts be? To which the answer is, that they are not inserted to distinguish the species, but in order to describe the family, and the total relations of the ovules and the stigmata of the family are better known by this general statement....

Though in a Natural group of objects a definition can no longer be of any use as a regulative principle, classes are not therefore left quite loose, without any certain standard or guide. The class is steadily fixed, though not precisely limited; it is given, though not circumscribed; it is determined, not by a boundary line without, but by a central point within; not by what it strictly excludes, but by what it eminently includes; by an example, not by a precept; in short, instead of a Definition we have a Type for our director.

Mill (1865) dropped the assumption of necessary and sufficient conditions, but he still believed that a closed-form definition was possible. He advocated a weaker criterion based on all of the necessary characteristics and a majority of the optional ones:
Whatever resembles the genus Rose more than it resembles any other genus, does so because it possesses a greater number of the characters of that genus, than of the characters of any other genus. Nor can there be the smallest difficulty in representing, by an enumeration of characters, the nature and degree of the resemblance which is strictly sufficient to include any object in the class. There are always some properties common to all things which are included. Others there often are, to which some things, which are nevertheless included, are exceptions. But the objects which are exceptions to one character are not exceptions to another: the resemblance which fails in some particulars must be made up for in others. The class, therefore, is constituted by the possession of all the characters which are universal, and most of those which admit of exceptions.

In his later philosophy, Wittgenstein (1953) repudiated his earlier logic-based approach by showing that ordinary words like *game* cannot be defined by necessary and sufficient conditions. Competition is present in ball games, but absent in solitaire or ring-around-the-rosy; organized sport follows strict rules, but spontaneous play does not; and serious games of life or war lack the aspects of leisure and enjoyment. Instead of differentiae that distinguish them from all other activities, games share a *family resemblance*: baseball is a game because it resembles the family of activities that people call games. Except for technical terms in mathematics, Wittgenstein maintained that most words are defined by family resemblances, and the meaning of a word is its use within a language. A word, he said, is like a chess piece, whose meaning is not determined by its physical shape, but by the rules for using it in the game.

All three of the nineteenth-century methods of definition have been refined and implemented in modern computer systems:

- **Logical.** A concept is defined by a genus or supertype and a set of necessary and sufficient conditions that differentiate it from other species of the same genus. This top-down method is still used in dictionaries and systems of deductive reasoning. It is the approach that Wittgenstein adopted in his early philosophy, but rejected in his later writings.
- **Fuzzy.** A concept is defined by zero or more necessary conditions and a preponderance of the optional ones ranked in order of importance. Variations of Mill’s criterion have been implemented in computational methods based on neural nets as well as *fuzzy sets*, *rough sets*, *cluster analysis*, and *latent semantic analysis* (LSA).
- **Prototype.** A concept is defined by an example or prototype, and any instance of the concept must resemble the prototype more closely than the prototypes of other concepts. This bottom-up method includes Wittgenstein’s family resemblances and the psychological prototypes studied by Rosch (1975). A computer implementation requires some measure of distance between prototypes, and the search for suitable measures has become an important research topic.

After two millennia of philosophical debate and 50 years of computer implementations, the modern consensus is not much different from Kant (1800):

> Since the synthesis of empirical concepts is not arbitrary but based on experience, and as such can never be complete (for in experience ever new characteristics of the concept can be discovered), empirical concepts cannot be defined. Thus only arbitrarily made concepts can be defined synthetically.
Such definitions... could also be called *declarations*, since in them one declares one’s thoughts or renders account of what one understands by a word. This is the case with *mathematicians*.

For any particular application, a top-down hierarchy of concepts can be legislated, but attempts to force all concepts into a universal, globally consistent hierarchy have failed.

### 3. From local features to global structures

Local features are useful for both top-down and bottom-up methods of classification, but the criteria for determining what features are significant depends more on purpose than on prominent, but superficial details. A rug and a sofa, for example, might be similar in color and texture, but not in size, shape, and function. A bird and a bat might be similar in flying ability, but in anatomical details a bat is more similar to a mouse than a bird. The criteria that determine what features to select and how to evaluate them depend on global considerations of structure, context, and purpose. The following is a summary of the kinds of criteria and the computational methods used or proposed for handling them:

- **Essence vs. accidents.** Plato and Aristotle recognized that some criteria are more fundamental than others. For example, the terms *rational animal* and *featherless biped* both distinguish the set of human beings. The first, they claimed, specifies the essence of what it means to be human, but the second depends on accidental properties. To distinguish essence from accident in biology, Aristotle proposed guidelines that are compatible with the classifications defined by DNA. Today, however, genetic engineering can insert genes from remote species into DNA and blur the distinction between essence and accident. In some theories, essence is represented by operators in modal logic, but the axioms or assumptions behind those operators determine the criteria that distinguish essence and accident.

- **Features or properties.** Monadic predicates, also called *properties*, serve as the differentiae in top-down definition and inheritance methods and in bottom-up fuzzy and prototype methods. Some features may be represented by monadic predicates, but others require dyadic predicates. Height and weight, for example, are numeric-valued features that would be represented by dyadic predicates, such as HasHeight(Tom,180 cm). Although they are as old as Aristotle, features are widely used in modern classification systems, including *description logics* [Baader et al. (2003)], *Formal Concept Analysis* [Ganter and Wille (1999)], and most statistical approaches.

- **External relations.** Many categories depend on relations with two or more arguments. Kinship terms, such as *mother* and *uncle*, are determined by logical combinations of dyadic predicates for biological and legal relationships; some terms, such as *ancestor*, even require recursive definitions. Other examples include employee, teacher, customer, or pilot. The relations link any individual of those categories to some external persons, places, things, or events. A man may become an uncle, for
example, without any knowledge of the fact. As an example of a category that requires four arguments, a person may become an heir \(x_1\) because of a will \(x_2\) that relates \(x_1\) to a deceased person \(x_3\) and a bequest \(x_4\). Individuals in these categories cannot be classified as such without considering information that may require lengthy search and inference.

- **Function or purpose.** The function of an artifact is its primary purpose, and a characteristic shape is incidental. Artifacts such as a chair, a spoon, or a dress have a recognizable shape because they are designed to fit the human anatomy. Others, such as a radio or a computer, have few restrictions on size or shape. Some things, such as toys and weapons, are defined only by function: any object, including a naturally occurring rock or an artifact designed for some other purpose, could be a toy if it is used in play, or a weapon if it is used to cause injury. Recognizing a toy or a weapon requires knowledge of context and purpose.

- **Spectrum or field.** Many phenomena vary continuously in one or more dimensions. The color spectrum, for example, varies smoothly from red to violet without sharp breaks, but all languages divide it with discrete words. Feelings can also vary continuously from happy to sad, but along any scale are clusters of words that cannot be linearly ordered: happy, joyful, and cheerful at one end; sad, gloomy, and melancholy at the other end; and OK, not bad, and so-so in the middle. Subtle distinctions among the words in any cluster are hard to define clearly, either for computers or for people.

- **Structure.** Although function is the ultimate goal of any artifact, many styles of design and construction have become traditional. A dwelling, for example, may be any structure that provides suitable living space and protection from the weather. But certain styles have become popular in various cultures: igloo, yurt, teepee, or Cape Cod salt box. Every culture can be identified by the design of its artifacts, and recognizing those designs is similar to recognizing natural kinds. Yet recognizing any structure, whether a pot or a hippopotamus, involves patterns of relationships defined by more complex combinations than a set of features. Methods of structural analysis include graph matching, pattern recognition, constraint satisfaction, and two- or even three-dimensional grammars.

- **Schema, Gestalt, frame, or chunk.** The problem of recognizing related structures in multiple manifestations was addressed by Kant (1787) with his notion of *schema*:

> Indeed, our pure sensible concepts are not based on images of objects, but on schemata. No image could ever adequately correspond to the concept of triangle in general, for it would never attain the universality of the concept which makes it applicable to all triangles, whether right-angled, acute-angled, or obtuse; it would always be limited to only one part of this domain... The concept of dog means a rule (Regel) according to which my imagination can form a general figure (Gestalt) of a four-footed animal, without being restricted to any particular figure given in experience or to any possible image I may draw in concreto. [Kant (1787), A. 141; 1800, B. 180]

Selz (1913, 1922) applied Kant’s term *schema* to networks of concepts used in perception and reasoning. Bartlett (1932) defined a schema as “an active organization of past
reactions or of past experiences” that operate as “a unitary mass.” In Gestalt psychology, Wertheimer (1925) addressed similar issues: “There are interdependencies (Zusammenhänge), in which the nature of the whole is not derived from the way the parts are put together, but rather the role of any part is determined by the inherent structural laws (Strukturgesetzen) of the whole.” In AI, the chunks by Newell and Simon (1972) and the frames by Minsky (1975) were inspired by the work of Selz, Bartlett, and the Gestalt psychologists.

- **Script, scenario, or procedure.** Most of the studies that characterize concepts by a schema or Gestalt have focused on static objects or patterns of objects. Actions and events also fall into regular patterns, which Schank and Abelson (1977) characterized by scripts. Businesses often specify procedures for routine tasks, and planners talk about scenarios they might encounter. Narayanan (1999) used Petri nets to characterize the patterns of actions and events in both narratives and business procedures. Sometimes a notation used for static structures can be adapted to dynamic scripts by replacing logical or spatial relations with temporal relations. Discourse representation structures [Kamp and Reyle (1993)] were originally designed to represent logical relations, but they were later extended to temporal relations; both versions can be translated to other notations for logic, such as predicate calculus and conceptual graphs.

- **Analogy.** Many categories are defined by global analogies, not by local features. As examples, stories classified by themes, such as jealousy, revenge, or coming of age, depend on analogous patterns that can be expressed in many different words. Similar patterns set in Elizabethan England, American frontier times, or modern business settings might have no words in common. In the analogy of the hydrogen atom to the solar system, the nucleus corresponds to the sun, the electron corresponds to the earth, and electrical attraction corresponds to gravitational attraction. This analogy depends on common global structures, not on any similarity in the local features. Analogical reasoning is a method of finding such patterns. For recognizing geometrical analogies on IQ tests, Evans (1963) ignored local differences to find similar global structures; Falkenhainer, Forbus and Gentner (1989) designed a domain-independent Structure Mapping Engine (SME) for recognizing analogies; and the VivoMind Analogy Engine (VAE) used more efficient global algorithms for finding mappable structures [Sowa and Majumdar (2003)].

- **Context, background knowledge, expectation.** The same patch of gray, when copied into two different photographs, might be interpreted as black or white. A puff of white noise, inserted in different sentences, might be interpreted as /p/, /t/, or /k/. The same word in different contexts might have totally different meanings or just slightly different variations of a common meaning. The same joke, told to the same people in different situations, might evoke laughter, silence, or anger. In discussing the need for background knowledge, Minsky (1968) proposed a number that has served as a goal for many projects:

> I therefore feel that a machine will quite critically need to acquire the order of a hundred thousand elements of knowledge in order to behave with reasonable sensibility in ordinary situations. A million,
if properly organized, should be enough for a very great intelligence. If my argument does not con-
vince you, multiply the figures by ten.

The Cyc project (Lenat, 1995) set out to implement Minsky’s goal of a million properly
organized knowledge elements. After 20 years of elapsed time and 650 person-years of
work in knowledge representation, Cyc has amassed 2 million axioms with 600,000 cat-
egories of entities organized in 6000 microtheories. In pure deduction, Cyc surpasses
the average human, but in finding relevant knowledge and learning new knowledge, it
cannot compete with a child. The caveat “if properly organized” is crucial: the proper
organization has not yet been discovered.

As this list shows, the criteria for recognizing, classifying, and interpreting things
and events are complex and varied. A feature-based classifier might be useful as the first
stage of perception, but context and purpose are essential for focusing attention on what
to observe, determining its relevance, and relating it to background knowledge. In
the sentence, “Yojo batted an eraser across the desk,” the words play and toy do not
occur. With the background knowledge that Yojo is a cat, cats are playful creatures,
and an eraser is a mouse-like object, one might interpret the action as playing and the
eraser as a toy. This issue is not limited to natural language understanding, since inter-
preting a movie or a photograph would require the same kind of analysis. Such inter-
pretations are necessary for reporting and classifying any observations. Luria (1968)
wrote a book about Shereshevskii, a man with a phenomenal memory for the exact
words and images he observed. Because of his memory, Shereshevskii got a job as a
newspaper reporter, but he was totally unsuited. His memory for detail was perfect, but
he could not interpret the detail, determine its relevance, or produce a meaningful sum-
mary. In effect, Shereshevskii behaved like a feature-based perceiver attached to a vast
database: he stored everything he saw or heard, but he could not interpret the relevance
of anything.

Every classifier, logical or fuzzy, feature-based or structural, depends on some model
for relating instances to categories. Figure 1, for example, illustrates the underlying
model for most neural nets, and Figure 2 illustrates the model for most logic-based sys-
tems. Data models incorporate general assumptions about patterns typically found in
well-behaved data; the popular bell-shaped curve, for example, is called “normal.”
Algorithmic models, such as neural nets, are based on assumptions about the mecha-
nisms that occur in some natural processes. Breiman (2001), a statistician who had
designed and used a wide range of models, emphasized the need for models that accu-
rately reflect the nature of the subject:

But when a model is fit to data to draw quantitative conclusions, the conclusions are about the model’s
mechanism, and not about nature’s mechanism. It follows that if the model is a poor emulation of
nature, the conclusions may be wrong. Approaching problems by looking for a data model imposes
an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statisti-
cal problems. The best available solution to a data problem might be a data model; then again it might
be an algorithmic model. The data and the problem guide the solution. To solve a wider range of data
problems, a larger set of tools is needed.
Every model is an approximation that extracts a simplified, computable mechanism from the overwhelming complexity of nature. As the statistician George Box observed, “All models are wrong, but some are useful.”

Features represent local information, which could be treated as independent variables, or they could be organized in a global pattern by a schema or Gestalt. In their original papers on chunks and on frames, Newell and Simon (1972) and Minsky (1975), respectively, tried to incorporate the full richness of the psychological models. In later implementations, however, the word frame was applied to data structures that do little more than package a few pointers. Those packages are useful, but they don’t model Selz’s (1913, 1922) schematic anticipation, Bartlett’s (1932) active organizations, or Wertheimer’s (1925) structural laws. Minsky (1987) continued to argue for a more globally organized society of mind, and Newell (1990) proposed a unified theory of cognition called “SOAR.” The pioneers in AI realized from the beginning that human intelligence depends on global mechanisms, but the challenge of bridging the gap between local features and global structure has not been met.

Chess playing illustrates the importance of the Gestalt effects. In applying Selz’s methods to chess, de Groot (1965) had chess players study positions and select a move while saying whatever came to mind, what moves or lines of play they were considering. He found no significant difference between masters and experts in the number of moves considered, depth of analysis, or time spent. The only significant difference was that the masters would usually focus on the best move at their first glance, while the nonmasters were distracted by moves that would dissipate their advantage. Former world champion Emanuel Lasker said that chess is a highly stereotyped game. Instead of exhaustively analyzing all options, the master simply looks at the board and “sees” which moves are worth considering.

After 40 years of chess programming, a computer was finally able to defeat the world champion, but only by a brute-force search. To discover what the human could see at a glance, the computer had to analyze the details of billions of chess positions. The Gestalt effect, which is significant on an 8 × 8 chess board with a maximum of 32 pieces, becomes even more pronounced in the oriental game of Go, which can have over 300 stones scattered around a 19 × 19 board. The number of possible patterns in Go is many orders of magnitude greater than in chess, and a brute-force search cannot overcome the human advantage of “seeing” the patterns. As a result, no program today can play Go beyond a novice level.

The schema or Gestalt theories appear in two different forms that are sometimes used in combination: discrete versions use patterns of graphs, and continuous versions are based on geometric models. The Gestalt psychologists emphasized image-like geometric patterns, but Selz’s graph patterns have been easier to implement. The term image schema [Lakoff and Johnson (1980), Oakley (2004)] has been used for geometric models mapped to patterns of discrete words expressed in natural languages. The psychological and linguistic evidence for such patterns is strong, and conceptual spaces [Gärdenfors (2000)] are a promising geometrical model that is designed to represent abstract concepts as well as physical images.
4. Categorization and reasoning

Categorization and reasoning are interdependent cognitive processes: every multistep reasoning process depends on categorization at each step; every one-step reasoning process is also a one-step categorization process, and vice versa; and every multistep categorization process involves multiple reasoning steps. As an example, consider the rule of deduction called modus ponens:

Premise: If P then Q.
Assertion: P.
Conclusion: Q.

This rule depends on the most basic technique of categorization: a comparison of two instances of A to determine whether they are identical. If the P in the assertion is not identical to the P in the premise, then a preliminary process of unification is required, which uses another technique of categorization: the specialization of both Ps to a common form. If the Ps are complex expressions, multiple categorization steps may be necessary.

Deduction is one of Peirce’s (1903) three methods of reasoning, each of which has a corresponding method of categorization. The following examples show how each of the three methods for reasoning about propositions has a corresponding method for dealing with categories. (For these examples, the words principle and fact are used as informal synonyms for proposition; a fact is considered more specialized than a principle, but either one could be arbitrarily complex.)

1. Deduction. Apply a general principle to infer some fact.

   Propositions: bird(x)→flies(x). bird(Tweety).
   Therefore, flies(Tweety).
   Categories: Birds⊂Flyers. Tweety∈Birds.
   Therefore, Tweety∈Flyers.

2. Induction. Assume a general principle that explains many facts.

   Propositions: bird(Tweety). bird(Polly). bird(Hooty).
   bat(Fred).
   flies(Tweety). flies(Polly). flies(Hooty). flies(Fred).
   Assume bird(x)→flies(x).
   Categories: Birds={Tweety, Polly, Hooty}. Bats={Fred}.
   Flyers={Tweety, Polly, Hooty, Fred}.
   Assume Birds⊂Flyers.

3. Abduction. Guess a new fact that implies some given fact.

   Propositions: bird(x)→flies(x). flies(Tweety).
   Guess bird(Tweety).
   Categories: Birds⊂Flyers. Tweety∈Flyers.
   Guess Tweety∈Birds.
In his pioneering work on symbolic logic, Boole (1854) used the same algebraic symbols for both propositions and categories. Since then, the notations have diverged, but the correspondences remain.

The methods of categorization and reasoning used in AI and related branches of computer science can be grouped in three broad areas, all of which have been under development since John McCarthy coined the term *artificial intelligence* in 1956. Most new developments may be considered incremental improvements, even though many of the “increments” are sufficiently radical to be important breakthroughs. Each of the three areas is characterized by its primary form of knowledge representation: features, structures, or rules.

1. **Feature analysis.** The oldest methods of both categorization and formal reasoning are based on collections of features, which may be processed by a variety of logical, statistical, and algorithmic techniques. The features could be monadic predicates, as in Aristotle’s differentiae, or they could be functions or dyadic predicates, in which the second argument (or *slot*) is a number or other *value*. A collection of features may be called a set, a list, a vector, a frame, or a logical conjunction. Among the feature-based methods are neural nets, decision trees, description logics, formal concept analysis, and a wide variety of vector-space methods, such as LSA.

2. **Structural analysis.** A collection of features, by themselves, cannot distinguish “blind Venetians” from “venetian blinds” or “dog bites man” from “man bites dog.” Instead of unordered collections, structural methods represent ordered dependencies and interconnections with directed graphs, including trees, forests, and nests of graphs within the nodes of other graphs. Such graphs are commonly used to represent the syntax and semantics of languages, both natural and artificial. Some versions represent rigorous logic-based formalisms, and others use informal heuristics. Among the structural methods are grammar-based parsers and pattern recognizers, constraint-satisfaction and path-finding systems, spreading-activation systems that pass messages through graphs, and graph-matching systems for finding patterns and patterns of patterns.

3. **Rule-based systems.** Rules are often used to process features or structures, but in a rule-based system, the rules themselves are the knowledge representation. The rules may be formulas in logic, or they may be less formal heuristic rules in an expert system. The more formal rule processors are called *theorem provers*, and the more informal ones are called *inference engines*. In general, the rules of a rule-based system may be used for multistep categorization in diagnostics and pattern recognition, or for multistep reasoning in planning and problem solving.

Large systems are often *hybrids* that combine more than one of these methods. A natural-language parser, for example, may use features for the syntactic and semantic aspects of words and build a *parse tree* to represent the grammatical structure of a sentence. A reasoning system may combine a *T-box* for terminology defined by a description logic with an *A-box* for assertions and rules.

The results of classification may be a static category or a dynamic control structure. Decision-tree systems, for example, are often used for robots because they learn quickly
and generate a decision tree that can be compiled into an efficient control mechanism. To train a robot, a teacher can take it “by the hand” and guide it through a task. At each step, the system records a vector of features that represents the current state of the robot together with the response that would take it to the next step. After that training pass, the system builds a decision tree that determines the response for any combination of features. If the robot makes a mistake on subsequent trials, the teacher can stop it, back it up to the point where the mistake occurred, and guide it through the next few steps. Then the system would modify the decision tree to prevent the robot from making the same mistake twice. It might, however, make similar mistakes in slightly changed conditions. Since the early days [Hunt (1962)], incremental improvements in the algorithms have enabled decision-tree systems to generalize more quickly and reduce the number of “similar mistakes.”

The boundaries between the three groups of AI systems are blurred because some features may be defined in terms of structures and some structures may be rule-like in their effect. As an example, a neural net for optical character recognition (OCR) might use a feature defined as “having an acute angle at the top.” That feature can be encoded in a single bit, derived from information about angles (geometrical structures) and is TopOf (a dyadic relation). As another example, analogical reasoning, which is based on structure mapping, can be used to derive the same kinds of conclusions as a rule-based system that uses induction to derive rules followed by deduction to apply the rules. That approach is the foundation for case-based reasoning [Riesbeck and Schank (1989)], but the principle was first stated by Ibn Taymiyya in his comparison of Aristotle’s logic to the analogies used in legal reasoning [Hallaq (1993)].

Ibn Taymiyya admitted that deduction in mathematics is certain. But in any empirical subject, universal propositions can only be derived by induction, and induction must be guided by the same principles of evidence and relevance used in analogy. Figure 3 illustrates his argument: Deduction proceeds from a theory containing universal propositions. But those propositions must have earlier been derived by induction with the same criteria used for analogy. The only difference is that induction produces a theory as intermediate result, which is then used in a subsequent process of deduction. By using analogy directly, legal reasoning dispenses with the intermediate theory and goes straight from cases to conclusion. If the theory and the analogy are based on the same evidence, they must lead to the same conclusions.

The question in Figure 3 asks for information Q about some case P. Answering that question by deduction requires a theory that contains some rule or combination of rules of the form \( \text{If } P \_ \text{ then } Q \_ \), where \( P \_ \) in the theory is a generalization of \( P \) in the given case. By a version of structure mapping called unification, \( P \_ \) in the theory is specialized to match \( P \). The same specialization, when applied \( Q \_ \), produces the answer \( Q \). If the question \( Q \) has one or more unknowns, as in Selz’s (1913,1922) method of schematic anticipation, the unknowns trigger a search to find the missing information. That search may take many steps, which may apply different rules of the same theory or even rules from different theories. But before any theory can be applied, it must have been derived by induction: analyze the evidence, apply a reliable methodology for
determining what commonalities indicate significant causal connections, and state those connections as “if-then” rules.

In analogical reasoning, the question Q leads to the same schematic anticipation, but instead of triggering the if-then rules of some theory, the unknown aspects of Q lead to the cases from which a theory could have been derived. The case that gives the best match to the given case P may be assumed as the best source of evidence for estimating the unknown aspects of Q; the other cases show possible alternatives. For each new case P, the same principles of evidence, relevance, and significance must be used. The same kinds of operations used in induction and deduction are used to relate the question Q to some corresponding part Q of the case P. The closer the agreement among the alternatives for Q, the stronger the evidence for the conclusion. In effect, the process of induction creates a “one-size-fits-all” theory, which can be used to solve many related problems by deduction. Case-based reasoning, however, is a method of bespoke tailoring for each problem, yet the operations of stitching propositions are the same for both types of reasoning.

Creating a new theory that covers multiple cases typically requires new categories in the type hierarchy. To characterize the effects of analogies and metaphors, Way (1991) proposed dynamic type hierarchies, in which two or more analogous cases are generalized to a more abstract type T that subsumes all of them. The new type T also subsumes other possibilities that may combine aspects of the original cases in novel arrangements. Sowa (2000) embedded the hierarchies in an infinite lattice of all possible theories. Some of the theories are highly specialized descriptions of just a single case, and others are very general. The most general theory at the top of the lattice contains only tautologies, which are true of everything. At the bottom is the contradictory or absurd theory, which is true of nothing. The theories are related by four operators: analogy and the belief revision operators of contraction, expansion, and revision [Alchourrón,Gärdenfors and Makinson (1985)]. These four operators define pathways through the lattice (Figure 4), which determine all possible ways of deriving new theories from old ones.
To illustrate the moves through the lattice, suppose that A is Newton’s theory of gravitation applied to the Earth revolving around the Sun, and F is Niels Bohr’s theory about an electron revolving around the nucleus of a hydrogen atom. The path from A to F is a step-by-step transformation of the old theory to the new one. The contraction step from A to B deletes the axioms for gravitation, and the expansion step from B to C adds the axioms for the electrical force. The result of both moves is the equivalent of a revision step from A to C, which substitutes electrical axioms for gravitational axioms. Unlike contraction and expansion, which move to nearby theories in the lattice, analogy jumps to a remote theory, such as C to E, by systematically renaming the types, relations, and individuals: the Earth is renamed the electron; the Sun is renamed the nucleus; and the solar system is renamed the atom. Finally, the revision step from E to F uses a contraction step to discard details about the Earth and Sun that have become irrelevant, followed by an expansion step to add new axioms for quantum effects. One revision step can replace two steps of contraction and expansion, but one analogy step can transfer a complete theory from one domain to another just by relabeling the symbols.

Different methods of walking through the lattice can produce the same results as induction, deduction, abduction, learning, case-based reasoning, or nonmonotonic reasoning. As a method of creative discovery, Fauconnier and Turner (2002) defined conceptual blending, which also corresponds to a walk through the lattice: given two cases A and B, derive a generic case T, which corresponds to the least common supertype in the lattice [as in Way’s (1991) method]; then use the expansion operator to specialize T to a blend B. Although all possible forms of reasoning, learning, and discovery can be reduced to walks in a lattice, the challenge is to find the correct path in an infinite lattice. In AI, search methods can be guided by a heuristic function, which estimates the distance from any given point to any desired goal. The term optimality, which Fauconnier and Turner adopted, is used in linguistics and psychology for the constraints that characterize a desirable goal; any computable definition of such constraints could be programmed as a heuristic function.
5. Levels of cognition

The methods of studying and modeling cognition differ from one branch of cognitive science to another, but all of them are abstractions from nature. As Breiman (2001) cautioned, any conclusions derived from a model are “about the model’s mechanism, and not about nature’s mechanism.” Furthermore, none of the models yet proposed in any branch of cognitive science can explain the full range of phenomena observed in humans and other animals in their natural environments. Figure 5 in the Separate Color Plate section shows some models that have been postulated for animals at various levels of evolution.

The cognitive systems of the animals at each level of Figure 5 build on and extend the capabilities of the earlier levels. The worms at the top level have rudimentary sensory and motor mechanisms connected by ganglia with a small number of neurons. A neural net that connects stimulus to response with just a few intermediate layers might be an adequate model. The fish brain is tiny compared to that of mammals, but it already has a complex structure that receives inputs from highly differentiated sensory mechanisms and sends outputs to just as differentiated muscular mechanisms, which support both delicate control and high-speed propulsion. Exactly how those mechanisms work is not known, but the neural evidence suggests a division into perceptual mechanisms for interpreting inputs, and motor mechanisms for controlling action. There must also be a significant amount of interaction between perceptual and motor mechanisms, and a simple neural net is probably an inadequate model.

At the next level, mammals have a cerebral cortex with distinct projection areas for each of the sensory and motor systems. If the fish brain is already capable of sophisticated interpretation and control, the larger cortex must add something more. Figure 5 labels it analogy and symbolizes it by a cat playing with a ball of yarn that serves as a mouse analog.
Whether the structure-mapping mechanisms of computer analogies can explain the rich range of mammalian behavior is not known, but whatever is happening must be at least as complex as a computer and probably much more so. The human level is illustrated by a typical human, Sherlock Holmes, whose is famous for his skills at induction, abduction, and deduction. Those reasoning skills may be characterized as specialized ways of using analogies, but they work seamlessly with the more primitive abilities.

Whatever the neural mechanisms may be, the functions of human intelligence operate at multiple levels. The instant reaction to touching a hot stove is controlled by a wormlike ganglion well before the brain perceives what happened. Although muscles can be controlled by conscious attention, they operate most efficiently when the details are left to the fishlike parts of the brain. In his subsumption architecture for mobile robots, Brooks (1986) distinguished eight levels of competence, each with increasingly more sophisticated kinds of processing:

1. Avoiding. Avoid contact with other objects, either moving or stationary.
2. Wandering. Wander around aimlessly without hitting things.
3. Exploring. Look for places in the world that seem reachable and head for them.
4. Mapping. Build a map of the environment and record the routes from one place to another.
5. Noticing. Recognize changes in the environment that require updates to the mental maps.
6. Reasoning. Identify objects, reason about them, and perform actions on them.
7. Planning. Formulate and execute plans that involve changing the environment in some desirable way.
8. Anticipating. Reason about the behavior of other objects, anticipate their actions, and modify plans accordingly.

Each of these levels depends on and subsumes the competence attained by the earlier levels. Each level responds to signals from the input sensors and generates output for the motor mechanisms. The first few levels by themselves could support an insectlike intelligence that responds directly without complex reasoning. For more sophisticated or intelligent behavior, the later levels may inhibit and take control from the earlier levels. But for danger signals, the primitive levels can override the inhibitions with a reflexive response.

The competence levels can be mapped to the six levels of the psyche defined by Aristotle (Hamlyn, 1993): nutrition, perception, desire, locomotion, imagery, and thought. These levels can help clarify the kinds of computation required for robots and their parallels to animal cognition. Nutrition and desire, for example, are two important levels that are missing in Brooks’ (1986) hierarchy.

1. Nutrition. For a mobile robot, nutrition corresponds the act of recharging its batteries from time to time. For a software agent, it would correspond to the procurement of computer time and space from some host.
2. Perception. For a robot, perception depends on input sensors and the ability to interpret the inputs. A television camera may provide a stream of data, but to see, the robot must convert the data to a representation of objects in the environment.
3. Desire. Aristotle’s general word for desire is *orexis*, which he further distinguished as appetite (*epithymia*), passion (*thymos*), and will (*boulēsis*). He classified appetite and passion as feelings shared with beasts and will as the result of rational thought. Moffat and Frijda (1995) made a similar distinction: the built-in equivalent of appetite or passion gives a robot a preference for certain kinds of states, but its will is determined by a logically derived plan for reaching a preferred state.

4. Locomotion. For mobile robots, locomotion includes the first three levels of Brook’s competence levels – avoiding, wandering, and exploring.

5. Imagery. Aristotle’s word *phantasia* means the appearance, presentation, or representation of images, either real or illusory. It would include Brooks’ levels of mapping and noticing.

6. Thought. Aristotle reserved the last level of the psyche for rational thought, which would include Brooks’ levels of reasoning, planning, and anticipating. Some aspects of those functions, however, are found in animals without speech. The psyche of an agent is its functional organization, and its level of sophistication depends on how much of the Aristotelian range of function it is able to support.

In AI systems, deduction has been developed in much greater depth than any other method of reasoning. Although methods of learning have been studied for years, no comparable level of success has been achieved. As an example, the following sentence was spoken by a child named Laura at 34 months of age [Limber (1973)]:

> When I was a little girl, I could go “geek, geek” like that; but now I can go “This is a chair.”

That single sentence incorporates a surprising amount of logical complexity: subordinate and coordinate clauses; earlier time contrasted with “now”; the modal auxiliaries *can* and *could*; the quotations “geek, geek” and “This is a chair”; metalanguage about her own linguistic abilities; contrast shown by *but*; and parallel stylistic structure. No program today can learn language at that level of complexity, and even some of the best language-generation programs have difficulty in producing sentences with such style.

Chomsky and his followers, such as Jackendoff (2002), believe that children learn language so rapidly because the human brain has an innate module that facilitates the acquisition of syntax. After studying human evolution, Deacon (1997) agreed that the brain has a predisposition to language, but that the evidence shows a slow evolution of the brain and vocal tract toward modern forms. It suggests that early hominids already had a rudimentary language, which gave individuals with larger brains and better speaking ability a competitive advantage. To analyze human and animal communication, Deacon used Peirce’s (1903) semiotic categories of *icon*, *index*, and *symbol* to classify the kinds of signs they could recognize and produce. He found that higher mammals easily recognize the first two kinds, icons and indexes, but only after lengthy training could a few talented chimpanzees learn to recognize symbolic expressions. Deacon concluded that if chimpanzees could make the semiotic transition from indexes to symbols, early hominids could. Once that transition had been made, language was possible, and the use of language would have promoted the co-evolution of both language and brain.
Although Jackendoff diverged from Chomsky by assuming graph-like conceptual structures, he replied to Deacon by repeating the claims for innate syntax. But if language depended on an innate syntactic module, a simulation of that module might enable computers to learn and understand language as easily as children. Yet syntax is not what makes language difficult for computers, and it is not essential for human understanding. People speaking a foreign language can make themselves understood even with badly fractured syntax, which is notoriously difficult for computers to parse. That supports Deacon’s hypothesis that symbols are more important than syntax: at first, children learn words as symbols for things; their predisposition to mimic adult speech enables them to learn patterns of words; and syntax results from learning to map speech patterns to cognitive patterns. Contrary to Chomsky, universal grammar is not the prerequisite, but the result of many generations of children preferring patterns that are easy to learn. Further research in semiotics may clarify these issues. Promising techniques include algebraic semiotics [Goguen (1999)], a dynamic logic of semiosis [Deacon (2004)], and a combination of Peirce’s semiotics with conceptual structures [Sowa (2000)].

In summary, computer cognition gives us a better appreciation for the full range of human abilities. The greatest challenge for cognitive science is to integrate thousands of separate studies into a working model of the brain. In AI, Newell (1990) and Minsky (1987) integrated 30 or more years of their own research into models of cognition that have shown some promise, but are no more successful overall than Lenat’s Cyc. Although AI is not yet able to simulate all aspects of human intelligence, it has contributed many valuable methods for enhancing intelligence and for testing hypotheses about intelligence.

References


Boole, G. (1854), An Investigation into the Laws of Thought (Dover Publications, New York).


Llull, Ramon (1303), Lògica Nova, vol. 4 of Nova Edició de les Obres de Ramon Llull, ed. by A. Bonner, Palma de Mallorca, 1998.


Selz, O. (1913), Über die Gesetze des geordneten Denkverlaufs (Spemann, Stuttgart).
Selz, O. (1922), Zur Psychologie des produktiven Denkens und des Irrtums (Friedrich Cohen, Bonn).
This page intentionally left blank
PART 2

SEMANTIC CATEGORIES
This page intentionally left blank
Chapter 7

SEMANTIC CATEGORIZATION

BRENDAN S. GILLON

McGill University

Contents

Abstract 167
1. Introduction 168
2. The notional approach to lexical categories 169
3. The notional approach to lexical subcategories 170
4. Structural approach to semantic categories 172
5. Coordinators and subordinators 175
6. English nouns 181
7. Conclusion 184
Acknowledgments 184
References 184

Abstract

The question addressed in this chapter is: what is a semantic category? The two views of linguistics – the structural and the notional – give rise to different answers. The chapter explains each view and illustrates how they differ with respect to the vexed question of the distinction between mass nouns and count nouns. Only the structural view has managed to provide useful and interesting answers to linguistic questions.
1. Introduction

The question to be addressed is: what is a semantic category? This question has two kinds of answers, one exoteric and one esoteric. If one takes the word *semantic* in its exoteric, or nontechnical, sense, then a semantic category is any category based on meaning. On this sense, any noun or noun phrase can be said to determine a semantic category. Racing cars and dark matter are both semantic categories, since they are categories based on the meaning of the noun phrases *racing cars* and *dark matter*. Semantic categories in this sense are not very interesting to linguists. What is interesting to linguists are semantic categories, when the word *semantic* is taken in its esoteric, or technical, sense. This is the notion of semantic category to be spelled out below.

Understanding what a semantic category is in the technical sense requires understanding what semantics is. There are two views of semantics, which correspond to two views of linguistics: the structural view and the notional view. Each of these views has a long tradition, reaching back over 2000 years. The notional view has its origins in classical Greece, namely in the theory of grammar dating back to Dionysius Thrax and beyond him to Aristotle and Plato. A notional view is one which seeks to define linguistic concepts in terms of nonlinguistic notions. The structural view is one which does not seek to reduce linguistic concepts to nonlinguistic notions. It has its origin in classical India, first clearly exhibited in the earliest extant grammar – a generative grammar in the modern sense of the term – Pāṇini’s Aṣṭādhyāyī, written around the 4th century BC. The structural approach slowly lost ground in India to the notional approach.

After lying dormant for more than 2000 years, the structural approach was reawakened by Leonard Bloomfield and developed by his successors Zellig Harris, Charles Hockett, Rulon Wells, and others. This structural approach was applied not only to phonology and morphology, as it had been by Pāṇini, but also to syntax. It remains at the heart of all formal approaches to linguistic theory.

In spite of the success of the structural approach, the notional approach remains strong, as seen in the work of cognitive linguists (for example, Gilles Fauconnier, George Lakoff, and Ronald Langacker), as well as in the work of the apparent heirs to the structuralist tradition, many of the self-styled generative linguists. Below, I shall set out what the structural approach to semantics is and show how it contrasts with the notional approach both to grammar in general and to semantics in particular.

The notional approach to the study of language is well illustrated by traditional grammar’s characterization of lexical categories. I shall first present traditional grammar’s characterization of English lexical categories and state some of the well-known criticisms of this approach. I shall then turn to the mass/count distinction, about which linguists of all stripes have voiced a view, and I shall show that the great majority of approaches, all of which are based on a notional approach to grammar, fail utterly to provide any account of the distinction. I shall show how greater progress can be made by relying on the structural approach. Next, I shall turn to a structural characterization of semantic categories, looking carefully at the distinction between subordinate and
coordinate conjunctions. I shall then return to the mass/count distinction to show how it is treated by a structuralist semantics.

2. The notional approach to lexical categories

Traditional English grammar identifies eight lexical categories, which are adaptations of the eight parts of speech enunciated in the earliest European grammar, that of Dionysius Thrax for classical Greek. Here is one formulation of the definitions of these categories for English:

- A **noun** is a word or a group of words that names a person, a place, an idea, or a thing (where things include objects, activities, qualities, and conditions).
- A **pronoun** is a word that stands for a noun. That for which a pronoun stands is its antecedent.
- A **verb** is a word or a group of words that expresses an action, a condition, or a state.
- An **adjective** is a word or group of words expressing a quality.
- An **adverb** is a word or group of words that modifies a verb, an adjective, or another adverb.
- A **preposition** is a word that shows a relation between its object and another word in the sentence.
- A **conjunction** is a word that connects two sentences or two parts of a sentence.
- An **interjection** is a word or group of words that expresses some feeling in an exclamatory fashion.

There is no doubt that the traditional, notional treatment of lexical categories tells us much of importance about English lexical categories. However, in the final analysis, it is inadequate. To begin with, one would expect the characterizations of the various lexical categories to be uniform. Some of the definitions characterize a part of speech in terms of its relation to other parts of speech. Thus, adverbs are characterized as modifying verbs, adjectives, and other adverbs. Others characterize a part of speech by appeal to substitutional equivalence, such as a pronoun being used to stand for a noun. In contrast, nouns, verbs, and adjectives are characterized by the kinds of entities they denote. These characterizations are notional ones, relying on ontic distinctions to distinguish among lexical items.

Defining lexical categories by what the members of the categories express is, in fact, not very helpful. Recall that a noun is a word or a group of words that names a person, a place, an idea, or a thing (where things include objects, activities, qualities, and conditions) and that a verb is a word or a group of words that expresses an event, a state, or a condition. Now consider the words *peace, war, famine, drought, armistice,* etc. Each of them denotes a state. Yet, they do not pattern with verbs, but rather with nouns. Words such as *picnic, game, ball, hurricane, fire, ceremony, lobotomy, crime, wake, meal,* and *rodeo* express events. But they too pattern with nouns, not verbs. An adjective is defined as expressing a quality. Consider the words *young, sad, silly, beautiful,* and *intelligent.* They denote qualities. But then so do the words *youth, sadness, silliness, beauty,* and *intelligence.* Yet they pattern with nouns, not adjectives.
3. The notional approach to lexical subcategories

Except for a handful of linguists, primarily cognitive linguists, the characterization of lexical categories by ontic differences has generally been abandoned. However, the use of ontic differences to distinguish subcategories of lexical categories remains widespread. The one we shall examine in some detail here is the division of common nouns into mass nouns and count nouns, a distinction which dates back at least to Otto Jespersen [(1909), vol. 2, ch. 5.2]. Jespersen noted that, while many English common nouns are felicitous both in the singular and in the plural, as shown below,

(1.1) Mary gave Jill a *flower*.
(1.2) Mary gave Jill *flowers*.

others are decidedly infelicitous in the plural.

(2.1) Mary gave Jill *dirt*.
(2.2) *Mary gave Jill *dirts*.


“A count noun designates a region that is specifically construed as being bounded within the scope of predication in a primary domain. By contrast, a mass noun designates a region that is not so construed.”

Langacker maintains that this contrast can be seen with respect to *brick* and *a brick*, as, for example, in the sentences in (3.1) and (3.3).

(3.1) Bill bought *brick*.
(3.2) Bill bought *bricks*.
(3.3) Bill bought *a brick*.

Note, however, that while *brick* in (3.1) is a mass noun and *bricks* in (3.2) is a count noun, no contrast of boundedness obtains between the sentence in (3.1) and the sentence in (3.2). Thus, boundedness is not a necessary condition for the distinction between mass nouns and count nouns. Nor is boundedness a sufficient condition, for the contrast of boundedness obtains between the noun *bricks* in (3.2) and the noun *brick* in (3.3), yet both are count nouns. Strikingly, Langacker says nothing about the possible contribution of the indefinite article to the contrast in construals.

Two other ontic distinctions have been invoked – namely, cumulativity of reference and divisivity of reference – in spite of the fact that these criteria have long been recognized as utterly inadequate. Quine (1960, p. 91) observed that, if a mass term such as *water* is true of each of two items, then it is true of the two items taken together. He
dubbed this semantic property of mass terms “cumulative reference.” This characterization, while apt, does not, however, distinguish mass nouns from count nouns, for cumulativity of reference also holds of plural count nouns. Just as it is the case that “if the animals in this camp are horses and the animals in that camp are horses, then the animals in the two camps are horses,” so it is the case that “if \(a\) is water and \(b\) is water then \(a\) and \(b\) together are water” [Link (1991)]. [See also Bunt (1985), p. 19.]

The third distinction, that of the divisivity of reference, suggested by Cheng (1973, p. 286–287), states that any part of something denoted by a mass noun is denoted by the same mass noun. The conjecture is that count nouns do not have this property, while mass nouns do. However, two classes of nouns show that this ontic distinction fails to distinguish the lexical classes of mass noun and count noun. On the one hand, a large class of words patterns morphosyntactically with mass nouns, yet their denotations have parts which do not fall within the same noun’s denotation. On the other, another large class of nouns patterns morphosyntactically with count nouns, yet their denotations have parts which also fall within the same noun’s denotation.

Over a decade before divisivity of reference gained any popularity as a necessary and sufficient condition by which to identify mass nouns, Quine (1960), p. 99), had pointed out that “…there are parts of water, sugar, and furniture too small to count as water, sugar, furniture.” In other words, not only does the mass noun furniture have parts which are not furniture, so too the mass nouns water and sugar have parts which are not water and sugar, respectively. Moreover, minimal pairs such as shown in Table 1 make it clear that divisivity of reference cannot be a necessary condition for a noun to be a mass noun.

It has been suggested that nouns such as drapery, furniture, pottery, jewelry, cutlery, silverware, hardware, gear, equipment, bedding, clothing, mail, toast, ammunition, and artillery are collective nouns. But this claim is groundless. The usual examples of collective nouns listed in English grammars – nouns such as team, herd, fleet, etc. – are

<table>
<thead>
<tr>
<th>Count nouns</th>
<th>Mass nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td>suitcases</td>
<td>luggage</td>
</tr>
<tr>
<td>suggestions</td>
<td>advice</td>
</tr>
<tr>
<td>shoes</td>
<td>footwear</td>
</tr>
<tr>
<td>undergarments</td>
<td>underwear</td>
</tr>
<tr>
<td>vehicle</td>
<td>traffic</td>
</tr>
<tr>
<td>guests</td>
<td>company</td>
</tr>
<tr>
<td>laughs</td>
<td>laughter</td>
</tr>
<tr>
<td>leaves</td>
<td>foliage</td>
</tr>
<tr>
<td>glasses</td>
<td>glassware</td>
</tr>
<tr>
<td>weapons</td>
<td>weaponry</td>
</tr>
<tr>
<td>animals</td>
<td>wildlife</td>
</tr>
<tr>
<td>cannons</td>
<td>artillery</td>
</tr>
</tbody>
</table>
count nouns. Moreover, collective nouns can take a prepositional phrase complement headed by the preposition of, which expresses the collective’s membership: a team of doctors, a herd of buffalo, a fleet of cars, etc. It is not possible to add such prepositional phrases to the mass nouns listed above.

Not only does the mass/count distinction not rely on collectivity, it does not rely on the distinction between the concrete and the abstract. It is certainly true that many abstract nouns are mass nouns: news, gratitude, fidelity, honesty, sincerity, chastity. But many abstract nouns are count nouns too: idea, mistake, universal, property.

The second class of nouns mentioned above includes the following: stone, rock, ash, string, cord, rope, and tile. These nouns behave both as mass nouns and as count nouns. But even in their guise as count nouns, they satisfy the criterion of the divisivity of reference. Suppose, for example, one has an 8-ft rope. Cutting it in half, one can be said to obtain two ropes of 4 ft each. And cutting each of these ropes in half again, one thereby obtains four ropes of 2 ft each.

In short, there is no ontic characterization of mass nouns. If the mass/count distinction cannot be given a satisfactory notional characterization, what satisfactory characterization can it be given? It can be given a structural characterization. To understand this characterization, let us see what a structural approach to semantics consists of.

4. Structural approach to semantic categories

Let me begin with what I call the distributional theory of meaning. As I stated above, the basic insights date back to Pāṇini. One idea is that a complex expression can be analyzed into minimal constituents. Another is that the sense of each minimal constituent in a sentence contributes to the sense of the entire sentence. As repeatedly observed by Sanskrit grammarians, one understands the sentences below differently:

(4.1) Horses are animals.
(4.2) Rocks are animals.

This difference in understanding is reflected in the fact that we judge the first sentence to be true, while we judge the second false. The difference in the truth value judgments is clearly attributable to the fact that where the word horse appears in one sentence, the word rock appears in the other, and these constituents are understood in different ways.

The next step came 2400 years later, when the American structuralist linguist Leonard Bloomfield (1933) rediscovered the context-sensitive rules used by Pāṇini in his grammar and applied them not only to morphology and phonology, as Pāṇini had done, but to syntax as well. Their use was subsequently greatly developed by Zellig Harris (1946), Rulon Wells (1947), Charles Hockett (1954), and others. The idea is that

1 The relevant rule is A. 1.2.45, that is, rule 45 of ch. 2 of Book I of his Aṣṭādhyāyī see Bronkhorst (1998) for a discussion].
each sentence can be analyzed into constituents of various kinds. These constituents of various kinds form substitution classes. In this view, lexical categories are the classes of the minimal constituents of the language.

A constituency analysis of a sentence identifies all the constituents of a sentence and the classes to which they belong. A constituent is indicated by a pair of brackets and the class by a label appended to the left bracket of the relevant pair. Here are examples of partial constituency analyses of two English sentences.

(5.1) The old man is walking with a cane.
(5.2) The old man [VP is walking [PP with a cane]].
(6.1) A man with a cane is sitting.
(6.2) [NP A man [PP with a cane]] [VP is sitting].

While Pāṇini and his successors understood that each sentence can be analyzed into minimal constituents and the sense of each minimal constituent contributes to the sense of the entire sentence, structuralist linguists understood that the contribution of the meaning of the minimal constituents to the meaning of the entire sentence they form is effected through the constituent structure of the sentence.

To see the truth and utility of these two assumptions, let us consider the following sentence:

(7) Galileo espied a patrician with a telescope.

Let the context for the utterance of the sentence be fixed – say, a speaker uttering the sentence in 1615. Let the circumstances with respect to which the sentence is to be evaluated be these: Galileo is looking out the window of his apartment in Venice through his telescope at a patrician walking empty-handed through Saint Mark’s Square. Is the sentence uttered in the fixed context true with respect to the stipulated circumstances? Yes and no. With respect to the circumstances envisaged, no patrician with a telescope is espied by Galileo, and so the sentence can be judged false. However, under the very same circumstances, a patrician has been espied by Galileo through his telescope, and so the sentence can be judged true. In this case, the variation in truth value judgment is due to the amphiboly of the sentence.

Whether the sentence is judged true or is judged false depends on which structure is assigned to it. Elementary constituency analysis of the sentence in (7) shows that it can be assigned either of two structures:

(7.1) Galileo [VP spotted [NP a patrician] [PP with a telescope]].
(7.2) Galileo [VP spotted [NP a patrician] [PP with a telescope]].

This observation – namely, that the sentence in (7) can both be judged true and be judged false with respect to one and the same set of circumstances – is explained on the basis of the two assumptions stated above, supplemented by what one might call Frege’s assumption – namely, the assumption that sense determines truth and reference.

Since the sentence in (7) can be judged both true and false with respect to the very same circumstances, one can conclude, by Frege’s assumption, that it has at least two senses. By the assumption of Pāṇini and of the structuralist linguists, the senses of the
sentence in (7) are determined by the combination of the senses of its minimal parts and its constituent structure. Therefore, either some minimal part has more than one sense or the sentence accommodates more than one syntactic structure. It can be shown, on the basis of elementary constituency analysis, that a proper contiguous part of the sentence in (7) – namely, saw a patrician with a telescope – accommodates two constituency analyses:

(8.1) [VP espied [NP a patrician] [PP with a telescope]].
(8.2) [VP espied [NP a patrician] [PP with a telescope]].

What is lacking from the foregoing is an explicit specification of how the values assigned to the parts determine the value of the whole.

The situation just described is analogous to the numerical expressions of the Indian numeral system. Here, a change in a numeral in a complex expression can change the value associated with the expressions of which it is a part. Thus, the difference in the values associated with the expressions 235 and 245 is attributable to the replacement of the digit 3 by the digit 4 and the difference in their values, namely, the number 3 and the number 4.

Let us look at this in more detail. The numerals used in arithmetic form an infinite set which can be rigorously specified in terms of a smaller set, indeed a finite set of ten elements, and two rules, one of which is recursive. The basic set comprises the following digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Let us call this set of symbols $D$. The set of numerals $IN$ is then, defined by two rules: (9.1) every digit is a numeral; and (9.2) every numeral put together with a digit is a numeral. The first rule is not recursive, but the second is. This definition can be stated more formally as follows:

(9.1) If $x \in D$, then $x \in IN$;
(9.2) If $x \in IN$ and $y \in D$, then $xy \in IN$;
(9.3) Nothing else is a member of $IN$.

To understand the definition in (9), let us see how the numeral 235 is obtained. By the first rule (9.1), the digit 2 is a numeral and so is the digit 3. So, by the second rule (9.2), 23 is a numeral. But, 5 is also a numeral (by the first rule). So, by the second rule, 235 is also a numeral. Note that this numeral is three digits long and it is obtained by two applications of the second rule, the first application to obtain the leftmost two digits 23, and then another application to obtain the leftmost three digits 235, that is, the entire numeral.

All numerals will be obtained in the same way. If a numeral contains $n$ digits, then it will result from $n - 1$ applications of the second rule, starting from the leftmost two digits and adding to the right, one after the other, the $n - 2$ remaining occurrences of digits.

While the rules in (9) specify an infinite set, they do not specify the values to be associated with that set. The basic idea, however, can be illustrated with the notation of modern arithmetic. Consider again the numeral 235. Recall that it was obtained by putting together the digits 2 and 3 to form the numeral 23, and that numeral was put together

---

2 This definition permits $IN$ to contain redundant numerals such as 00129. We tolerate these superfluous numerals because to ban them would involve distracting complications and their presence does no harm.
with the digit 5 to form the numeral 235. To assign a value to the numeral, we proceed as follows. First, assign to each digit its usual value. Thus, the digit 2 is assigned two, the digit 3 three, and the digit 5 five. Next, assign values to each numeral obtained in the construction of 235. To assign a value to 23, one multiplies the value assigned to the digit 2, namely two, by ten and adds to that the value assigned to the digit 3, namely three, thereby obtaining a value of 23. We can apply the same procedure to obtain the value of 235, namely, to multiply the value of the numeral 23 by ten, to obtain 230, and to add to it the value of the digit 5, namely five, thereby obtaining 235.

The value of each numeral is obtained in the same way. If a numeral contains \( n \) digits, then a value is assigned to the numeral corresponding to the leftmost two digits of the original numeral. This same procedure is applied \( n - 2 \) times to obtain values for each of the \( n - 2 \) numerals. The assignment of values is recursive and it follows the recursive structure of the numerals.

Here is the formal definition of the recursive assignment of values to the numerals in \( IN \).

\[
\begin{align*}
(10.1) & \quad \text{If } x \in D, \text{ then } i(x) = x; \\
(10.2) & \quad \text{If } xy \in IN \text{ and } y \in D, \text{ then } i(xy) = i(x) \times 10 + i(y)
\end{align*}
\]

(where bolded \( x \) is the natural number usually corresponding to the symbol \( x \)).

The rules in (10) show how a change in a digit in the numeral 235 leads to a change in its value. The replacement of 3 in 235 by 4 results in 245, whose value is 245. The rules in (10), applied to the numeral 245, yield precisely the value of 245. In short, the rules in (9) and (10) permit us to see how our understanding of a complex numeral changes with a change in one of its constituents.

An idea of how one might account for the way in which values associated with constituent expressions contribute to the value associated with the expression they constitute comes from logic, where sets become surrogates for meaning and truth values for propositions. We now provide an illustration of this application, which shows how parts of logic can prove elucidating and yet not trivial.

5. Coordinators and subordinators

Just as the connectives of propositional logic are used to form complex formulae from either propositional variables or formulae, so English connectors such as and, or, if, and because are used to form complex declarative sentences from either simple or complex declarative sentences. It is natural to explain our judgments of the truth conditions of a compound declarative sentence in terms of the truth conditions of its constituent declarative clauses and a truth function hypothesized as characterizing at least part of the meaning of the connector, in the same way as the truth value of a formula is determined by the truth value of its constituent formulae and the truth function of its main connective.

By way of an example, consider the compound declarative sentence \( \text{It is raining and it is cold} \). It seems to be the case that we judge the entire sentence to be true when and only when we judge each of its constituent declarative clauses to be true (Table 2).
But this is exactly the truth function \( o \), which interprets the propositional connective \( \land \). It seems natural, then, to hypothesize that the truth function \( o \) characterizes at least part of the meaning of the English connector \( \text{and} \).

Just as one cannot obtain a value for a complex piece of notation on the basis of an assignment of values to its basic symbols without knowing how the notation is structured, so one cannot obtain an interpretation of a compound declarative sentence in a natural language like English on the basis of interpretations of its constituent clauses without knowing how the constituent clauses combine. For this reason, we must ascertain how the connectors combine with other constituent clauses to form compound clauses or sentences. We start below, then, with a review of the principal syntactic facts pertaining to English connectors and their clausal syntax.

English, like other languages, distinguishes between coordinated and subordinated clauses and between coordinators (for example, \( \text{and} \) and \( \text{or} \)) and subordinators (for example, \( \text{if} \) and \( \text{because} \)). Here are the principal differences. [Quirk et al. (1985), ch. 13.5-13.21] First, coordinators may connect verb phrases, whereas subordinators may not.

(11.1) Dan [VP drank his coffee] and [VP left quickly].
(11.2) * Dan [VP did not drink his coffee] because [VP left quickly].

Another difference manifests itself in verb ellipsis, or what is known in the linguistic literature as gapping. Roughly, when two English clauses that have parallel syntactic structure and share the same verb are connected by a coordinator, the verb of the second clause may be omitted. This is not the case when the connector is a subordinator.

(12.1) [5 Bill encouraged Mary] and [5 Dan encouraged Maggie].
(12.2) [5 Bill encouraged Mary] and [5 Dan _ Maggie].
(13.1) Bill encouraged Mary, because Dan encouraged Maggie.
(13.2) *Bill encouraged Mary, because Dan _ Maggie.

Third, coordinators never occur one immediately after another. However, subordinators may.

(14.1) *John is unhappy and but he does what he is told.
(14.2) Bill left, because if he hadn’t, he would have been in trouble.
(14.3) We don’t need to worry about Carol, because if when she arrives we are not home, she can let herself in with the key I lent her.

<table>
<thead>
<tr>
<th>It is raining</th>
<th>It is cold</th>
<th>It is raining and it is cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>T</td>
<td>F</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Table 2
Truth function for the English coordinator \( \text{and} \)
Moreover, a coordinator may immediately precede a subordinator.

(15) Dan asked to be transferred, because he was unhappy and because he saw no chance of promotion.

Finally, the English subordinator may occur at the beginning of the pair of connected clauses, whereas the English coordinator may not.

(16.1) [S It is cold] and [S it is raining].
(16.2) [S It is cold] because [S it is raining].
(17.1) * And [S it is raining] [S it is cold].
(17.2) Because [it is raining] [it is cold].

It is clear from the preceding that coordinated clauses consist of two clauses, one on either side of a coordinator. This constituent analysis is nicely depicted by a constituency tree in Figure 1, commonly known as a phrase marker (where $S$ is a clause and $C_c$ is a coordinator). It is convenient to specify such a structure with a constituency rule (better known as a phrase structure rule): $S \Rightarrow S C_c S$. This rule says that two clauses flanking a coordinator form a clause.

Similar trees can be used to analyze the structure of the formulae of propositional logic. In such a tree, known as a categorematic construction tree, the propositional connective $\land$ occupies a position similar to that occupied by the coordinator $\text{and}$ in a constituency tree. To see this, consider, for example, the categorematic construction tree below (left) for the formula schema $\alpha \land \beta$. Replace each formula in it with the name of the set to which it belongs – namely $FM$ (for formulae) – and the propositional connective by the name of the set to which it belongs – namely $BC$ (for binary connective). This yields the middle tree in Figure 2. Next, replace the label $FM$ with the label $S$ and the label $BC$ with the label $C_c$. There is, however, also a disanalogy. Nodes in a categorematic construction tree are labeled by formula, with the bottom nodes being labeled by atomic formulae. Nodes in a constituency tree are labeled by syntactic categories and its terminal nodes are labeled by expressions of English. This means that the constituency tree for a pair of coordinated English clauses, shown in Figure 3, contains more structure than what is shown in Figure 1.

The English coordinator $\text{and}$ bears other striking similarities to the propositional connective $\land$. First, the truth of a clause which is formed from two independent clauses coordinated by $\text{and}$ depends on the truth of its constituent clauses in the same way as the

\[ S \Rightarrow S C_c S. \]

Fig. 1. Constituency tree for coordination.

Ch. 7: Semantic Categorization

177
truth of a complex formula formed from two formulae connected by $\land$. Consider, for example, the two simple independent clauses: it is raining and it is cold. It is indisputable that the clause formed from these two independent clauses with the coordinator $and$ is true under the circumstances in which both coordinated clauses are true, and not otherwise. And, as we saw above, this is exactly the truth function that interprets $\land$ of propositional logic. This truth function enjoys several logical properties which it seems to share with the coordinator $and$. To begin with, both are commutative.

(18.1) It is raining and it is cold.
(18.2) It is cold and it is raining.
(19.1) Mary studies at McGill and John studies at Concordia.
(19.2) John studies at Concordia and Mary studies at McGill.

In addition, just as one can infer either of the immediate subformulae of a formula whose main connective is $\land$, so can one infer either independent clause of a pair of independent clauses forming a clause connected by the coordinator $and$. Thus, in propositional logic, the following hold:

(20.1) $p \land q \models p$
(20.2) $p \land q \models q$
(20.3) $p, q \models p \land q$
The analogues of these entailment relations hold in English as well$^3$.

(21.1) it is raining and it is cold ENTAILS it is raining.
(21.2) it is raining and it is cold ENTAILS it is cold.
(21.3) it is raining, it is cold ENTAILS it is raining and it is cold.

It should be evident that the same rule of interpretation which applies to the cate-
gorematic construction tree to assign a value to the mother node on the basis of values
assigned to the daughter nodes, can apply to the corresponding part of the constituency
tree above, so that in both cases the top mother node acquires a value determined by the
values assigned to its daughter nodes, as shown below (Figure 4). The constituency rule
$S \Rightarrow S_1 C_c S_2$ characterizes the structure of the constituent. A corresponding interpreta-
tion rule must be stated to say how the constituent is to be interpreted on the basis of
the interpretation of its immediate constituents. The rule is quite simple: $v(S) = v(C_c)$
($v(S_j)$, $v(S_2)$).

However, the question remains: how are values assigned to the nodes labeled $C_c$ and
and? As we saw above, the English coordinator and includes in its meaning the truth func-
tion $o$. Thus, and $\rightarrow o$. (that is, $v$ (and) $= o$). Finally, there must be a rule which states
how the value of the node labeled with and is related to the value of the node labeled with
$C_c$. Now, and is a lexical item and $C_c$ is a lexical category. The rule is that the value of
the lexical category $\kappa_\lambda$ of a lexical item $\lambda$ is the same as the value of the lexical item itself (that
is, $\kappa_\lambda = v$ iff $\rightarrow - v$). The entire constituency tree, fully interpreted, is given in Figure 5.

In addition, then, to the constituency rule $S \Rightarrow S_1 C_c S_2$ and the lexical rule $C_c \Rightarrow$ and,
there are two rules of interpretation, which are stated below.

IR 0:
Let $[X Y]$ be the syntactic structure. Let $v$ be an interpretation function.
Then, $v(X) = v(Y)$.

IR 1:
Let $[S S_1 C_c S_2]$ be the syntactic structure. Let $v$ be an interpretation function. Then, $v(S) = v(C_c)$ ($v(S_j)$, $v(S_2)$).

$$
\begin{align*}
\alpha \land \beta & \mapsto \sigma_\land(v_1, v_2) \\
S & \mapsto \sigma(v_1, v_2) \\
\alpha & \mapsto v_1 \\
\land & \mapsto \sigma_\land \\
\beta & \mapsto v_2 \\
S & \mapsto v_1 \\
C_c & \mapsto \sigma_\land \\
S & \mapsto v_2
\end{align*}
$$

Fig. 4. Interpreted categorematic construction tree and interpreted constituency tree.

$^3$ '$\models$' is a formally defined relation which holds between a set of formulae of classical propositional logic
and a single one of its formulae. It is not defined for sentences of English. The analogous relation for declar-
ative sentences of English is referred to by ENTAILS.
But the adaptation of logic to language is not always quite so straightforward. Consider now the English subordinator *if* and the propositional connective $\rightarrow$. As observed above, subordinated clauses have two structures, both of which differ from coordinated clauses (Figure 6).

These constituency trees can be specified by three rules. The first – $D \Rightarrow C_s S$ – states that a clause ($S$) preceded by a subordinator ($C_s$) constitutes a dependent clause ($D$). The second two – $S \Rightarrow D S$ and $S \Rightarrow S D$ – state that a clause ($S$) either preceded or succeeded by a dependent clause ($D$) constitutes a clause ($S$). Below, we list the constituency rules given so far:

\begin{align*}
CR 1: & \quad S \Rightarrow S C_c S \\
CR 2: & \quad S \Rightarrow S D \\
CR 3: & \quad S \Rightarrow D S \\
CR 4: & \quad D \Rightarrow C_s S
\end{align*}

(where $C_c$ is the lexical category of coordinators and $C_s$ is the lexical category of subordinators.)

Direct inspection of the constituency trees given above reveals that coordinated clauses and subordinated clauses have different structures. Moreover, we saw that the structure of the categorematic construction tree for a formula whose main connective is a binary one is the same as that of a coordinated clause. It follows, as we shall see, that a truth function which interprets a binary connective cannot be used to interpret a subordinator.

The truth function $o_\land$ simply cannot apply in either of the syntactic structures appropriate to subordinating conjunctions. The reason is clear: whereas the coordinator has two sister clauses, the values of which provide a pair of values on which the binary truth function interpreting the coordinator can operate, the subordinator has only one sister clause, which provides but one value for the binary truth function to operate on. But the truth function $o_\land$ requires two truth values.

The solution is to find a truth function which is equivalent to $o_\land$ but can apply in the syntactic structures of subordinate clauses. The following shows how this might be
done. Suppose that the apodosis is assigned $T$ and the protasis is assigned $T^4$. We know that the top node must be assigned $T$. The questions are: what is to be assigned to the node $D$ and what is to be assigned to the node $C_s$? $D$ must be assigned a unary truth function. Moreover, since its sister node is assigned $T$ and its mother node is also assigned $T$, it must map $T$ to $T$. Suppose, next, that $D$’s sister node is assigned $F$. In that case, the mother node must be assigned $F$, for any conditional sentence with a true protasis and a false apodosis is false. In short, then, if the protasis clause is assigned $T$, then the $D$ node must be assigned the unary truth function which maps $T$ to $T$ and $F$ to $F$. This is known as an identity function. We shall designate it as $o_i$.

Next, suppose that the protasis node is assigned $F$. We know, then, that the top node must be assigned $T$, no matter what is assigned to the apodosis. This implies that the $D$ node must be assigned the constant unary truth function which maps both $T$ and $F$ to $T$. We shall call this function $o_T$.

In short, then, if is to be interpreted by a function which maps $T$ to the $o_i$ and $F$ to $o_T$. The reader should check that this interpretation of if works, regardless of whether the subordinate clause precedes or succeeds the main clause.

What we have seen in the foregoing is this: first a determination of the distribution of clauses and their connectors. Investigation of the constituency as well as the linear order of the constituents led to the formulation of constituency rules furnishing the relevant constituency. Paired with the constituency rules are interpretation rules which show how the truth values of constituent clauses determine the truth value of the clause they comprise. Finally, suitable meanings were assigned to the coordinators and subordinators. The overall idea can be pictured as in Figure 7.

6. English nouns

Having seen how to proceed in the case of coordinators, subordinators, and the compound sentences they help to form, let us return to the problem posed by the mass nouns

---

4 The protasis and apodosis are the subordinate and superordinate clauses of a conditional sentence, what philosophers often call the antecedent and the consequent, respectively.
and count nouns. The place to begin is with a characterization of the distributional properties of English nouns. English nouns can be partitioned into four classes, according to whether or not the nouns in question tolerate or resist being preceded by a determiner and whether or not they admit the contrast between singular and plural.

English proper nouns such as *Alice, Duns Scotus, Canada, and Paris* are not freely preceded by determiners. Some proper nouns such as *The Andes* and *The Hague* have the definite article as a component. But notice that the definite article cannot be felicitously replaced by any other determiner nor can it be preceded by any determiner. Moreover, proper nouns do not show a singular/plural contrast. In general, proper nouns occur only in the singular. Those which do occur in the plural never occur in the singular.

English pronouns include not only the well-known personal pronouns, and the interrogative and relative pronouns, but also what used to be called indefinite pronouns, for example, *someone, each, none, all*, etc. Unlike English proper nouns, these words have both singular and plural forms, though in many cases one is a suppletion of the other. Like English proper nouns, however, they do not permit determiners to precede them.

English count nouns are like English pronouns insofar as each English count noun has both a singular form and a plural form. Yet, unlike English pronouns, they are freely preceded by determiners.

English Count Nouns: *table, man, wire, pipe, machine, carrot, friend, belief, pebble, ocean, leaf, letter*, etc.

English mass nouns are also freely preceded by determiners, but unlike English count nouns, they do not show a singular/plural contrast.
English Mass Nouns:
water, oil, gold, wiring, water, piping, equipment, broccoli, furniture, knowledge, metal, wine, information, foliage, mail, etc.

The foregoing observations are conveniently summarized in Table 3.

As was just noted, English count nouns admit a morphological contrast between singular and plural; mass nouns do not, being almost always singular. Correlated with this are several other generalizations. Cardinal numerals and quasicardinal numerals (e.g., several) modify count nouns, never mass nouns. For example, two drinks is acceptable, since drink is a count noun; but two milks is not, since milk is a mass noun. Moreover, little and much modify mass nouns, never count nouns; whereas few and many modify count nouns, never mass nouns. Mass nouns do not tolerate the indefinite article (*an advice), whereas singular count nouns do (a suggestion). The pronoun one may serve as the antecedent of count nouns, but not of mass nouns.

(22.1) Mary gave Jill advice and John gave her some/*one too.
(22.2) Mary gave Jill a suggestion and John gave her one too.

Again, these generalizations are tabulated in Table 4 below.

While mass nouns and count nouns are clearly distinguishable by distributional criteria, as just shown, they are not distinguishable by any ontic characterization such as boundedness, cumulativity, or divisivity, as established in Section 3. What, if any, semantic difference can be correlated with the mass/count distinction? It is nothing more than this: count nouns are marked, or specified, for having an atomic denotation, while mass nouns are unmarked, or unspecified, for whether or not their denotation is atomic.

That the mass/count distinction grammaticalizes nothing more than the requirement to have an atomic denotation, or the lack of such a requirement, is evidenced by the large number of English common nouns which participate in both lexical categories; these nouns have denotations that, according to common sense, are ontologically quite diverse. Count nouns denoting trees, such as oak, double as mass nouns for the wood of such trees. Count nouns for animals that humans eat, such as lamb, double as mass nouns for the edible flesh of such animals. Similarly, many nouns for edible vegetables, such as potato, double as nouns for the food obtained from such vegetables. At the same time, many mass nouns double as count nouns, acquiring thereby the requirement that the denotation be atomic. What those atoms are depends on the individual denotations.

<table>
<thead>
<tr>
<th></th>
<th>Occurs with a Determiner</th>
<th>Admits the contrast of singular and plural</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Proper noun</strong></td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td><strong>Pronoun</strong></td>
<td>–</td>
<td>+</td>
</tr>
<tr>
<td><strong>Mass noun</strong></td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td><strong>Count noun</strong></td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
In the case of food, the atoms are units of packaging or service: pizza, hamburger, chocolate, candy, cake, etc. In the case of certain artifacts or naturally occurring objects, atoms may be determined by perceptual properties: rope, string, stone, tile, ash, shadow, etc. In other cases, the atoms may simply be instances of what the mass noun denotes: noise, detail, effort, thought, pain, virtue, etc.

This way of treating the mass/count distinction is compatible with the view of semantics set out in Section 4. The details are found in Gillon (1992, 1999).

7. Conclusion

The question which was raised at the beginning of this Chapter is: what is a semantic category? The answer is that semantic categories are those categories required by linguistic theory to explain how the meaning of a complex expression is determined by the meaning of the expressions constituting it. To ascertain these categories, one must first ascertain the structure, that is, the syntactic structure, which provides the part/whole relationship on which the determination of meaning depends. Semantic categories are not, in any interesting sense, the categories which can be expressed by the expressions of a language.

Acknowledgments

This paper was presented on October 31, 2003, at the Linguistische Kategorien (findung), held at the Ludwig–Maximilians–Universität (Munich, Germany). Research leading up to this paper has benefited from participation in the Mental Lexicon Project (SSHRC 412-2001-1009).

References


Ojeda, A. (1993), Linguistic Individuals (Center for the Study of Language and Information, Stanford, CA)

CLSI Lecture Notes no. 31.


Chapter 8

EMOTION CATEGORIES ACROSS LANGUAGES*

JAMES S. BOSTER
University of Connecticut

Contents

Abstract 188
1. Introduction 189
2. Methods of assessing cultural emotion systems 191
   2.1. The Method of Translation 191
   2.2. The Method of Mapping 198
3. Theories of emotion 210
4. Cross-cultural scenarios as a tool to compare emotion categories 213
5. Conclusion 218
References 220

* A portion of the research presented here was supported by grants from the National Science Foundation (9910445 and 9910465) and the University of Connecticut. I thank Roy D’Andrade, Cornelia Dayton, Phoebe Ellsworth, Asifa Majid, Kateryna Maltseva, John Shaver, and Richard Shweder for their comments and suggestions. I thank Claire Lefebvre and Henri Cohen for their hard work in putting this volume together and Zofia Laubitz for her assiduous copyediting. I also thank the Waorani, Shuar, Americans, Poles, Mexicans, and Italians who have patiently responded to my curiosity about how they see their emotional worlds.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved

187
Abstract

In this chapter, I assess the degree to which emotions are biologically endowed or culturally constructed, explore the nature of the similarity and difference among cultural emotion systems, and compare methods of carrying out that exploration. Rejecting both extreme universalism and relativism, I insist that emotions are at once absolutely biologically endowed and completely locally culturally constructed. I distinguish two major strategies for investigating the similarities and differences among cultural emotion systems. In the “translation method,” one proceeds by finding equivalent sets of emotion terms in two languages, asking native speakers to judge the similarity of the meanings of the terms, and assessing the degree to which the semantic relationships among the terms are similar for the two languages. In this case, translation is a prerequisite of the analysis. Unfortunately, this method presumes the answer to the research question itself. The degree of correspondence between emotion systems cannot be assessed if one presumes to know the corresponding terms from the outset. In the alternative “mapping method,” one proceeds by presenting emotionally evocative stimuli to native speakers, asking them to identify the emotion expressed in the emotional stimuli, and assessing the degree to which the implied similarity of the stimuli is similar for the two languages. In this method, translation is a consequence rather than a prerequisite of the analysis. I present the results of employing the mapping method to study the similarities and differences in the naming of emotional states from facial expressions and from emotional scenarios across a number of languages. I close with an outline of a research design for continued exploration of the patterns of similarity and difference among cultural emotion systems.
1. Introduction

This chapter addresses three questions. First, to what degree are emotions biologically endowed or culturally constructed? Second, what is the nature of the similarity and differences among cultural emotion systems? Third, what does one learn from different ways of comparing cultural emotion systems? Most of the chapter concentrates on answering the third question, both by reviewing the research designs and conclusions drawn by other investigators and by exploring the implications of my own research. I will close with an outline of a method of exploring cultural emotion systems that responds to the criticisms that arise from the review of earlier research.

Concern with the first question has dominated the literature on culture and emotions, and, as with many arenas of debate, there is a continuum of positions. At one end are universalists who argue that basic human emotions are experienced and expressed largely independent of the individual’s culture background [e.g., Ekman (1972), Izard and Buechler (1980)]. At the other end are relativists, who argue that individuals from different cultures inhabit incommensurate emotional worlds. Close to this pole are Rosaldo (1980, 1984) and Lutz (1985, 1987, 1988). In the middle are those who distinguish between the inner experience of subjective emotional states (which are potentially universal) and the conceptual systems by which the emotions are defined and classified (which are necessarily culturally specific) [e.g., Gerber (1985)].

1 It is difficult to find anyone in contemporary literature willing to take as radical a stance as Whorf (1941) or Lévy-Bruhl (1910/1985) and claim that different peoples inhabit incommensurate emotional and conceptual worlds. That is why I place Rosaldo and Lutz close to, but not at, the relativist pole. Rosaldo backs off from a radical cultural relativism or determinism [cf. Spiro (1984), pp. 330, 345–346] and acknowledges (1984, p. 142) that it is difficult to insist that “there is nothing universal about such things as happiness and anger.” Nevertheless, she firmly asserts (1984, p. 142) “the Balinese no more feel ‘guilt’ than we feel lek, the Balinese emotion closest to our ‘shame.’” Similarly, Lutz is careful not to deny the possibility that there may be some ethnopsychology tenets that are universally held (1985, pp. 69–70) and affirms that her own ethnography is “both implicitly and explicitly comparative” (1985, p. 68). If ethnography can be comparative, then obviously the emotion worlds inhabited cannot be incommensurate. However, she is adamant about the chasm separating Western and other ethnopsychologies (1985, p. 38–39):

Academic psychology has taken English emotion words (such as ‘fear,’ ‘love,’ ‘anger,’ and ‘disgust’), has reified what are essentially American ethnopsychological concepts, and has accepted them, often unquestioned, as the conceptual apparatus of scientific inquiry. Given this limited cultural base, it would be surprising if the emotions, exactly as distinguished, conceptualized, and experienced in American society, emerge as universals. Exact this has been assumed, however, and then ‘proved’ by Western researchers [Ekman (1974), Sorenson (1976)]. While it has been considered of great importance to ascertain whether some non-Western peoples ‘feel guilt’, the question does not arise as to whether Americans experience the New Guinea Hageners’ emotion of *popol* ‘outrage over the failure of others to recognize one’s claims’ [Strathern (1968)] or whether they are deficient in the ability to experience the Ifaluk emotion of *fago* ‘compassion/love/sadness.’

2 Shweder and Bourne (1984, pp. 158–159) recognize another stance, “confusionism,” that may not lie on the universalism–relativism continuum:

Confusion(ism) calls for the honest confession that one fails to comprehend the ideas of another. … We would, however, like to confess, right here, that not infrequently we are left in a muddled condition,
Most reasonable investigators of culture and emotion choose the middle position, if only for the pragmatic reason that it is the one that permits a meaningful empirical enterprise of comparing how emotions are experienced and interpreted in different cultural settings. If all emotions are completely universal, there is nothing to compare; if all emotions are entirely culturally specific, there is no basis for comparison. Thus, the only position that allows investigators to engage in detailed comparison of emotion systems is the middle ground.

However, rather than being just a useful but wimpy compromise between the clear-headed at either end, the middle ground is also the only sensible position, because the biology–culture dichotomy is a false one. Emotions are at once everywhere absolutely biologically endowed and everywhere completely locally culturally constructed. By virtue of being members of the same species, humans share the same physiological infrastructure of hormones and neurons, viscera and sinews that embody the human capacity for emotional response. But what counts as an occasion to feel one emotion or another is always defined in locally culturally specific terms. Furthermore, the differences between the extreme positions may be more rhetorical than real. Thus, for example, Ekman and fellow universalists are willing to acknowledge a role for culture, but they assign that role to emotional display rules. In this way, they are able to have their cake and eat it, keeping the emotions themselves universal and allowing culture to play a role in setting the conditions for how and when those universal emotions are experienced and expressed. Similarly, despite the fact that the relativists envision vast chasms between emotional worlds, somehow they are able to vault these chasms without difficulty and come back with reports of other cultural meaning systems.

I believe it is also fair to say that the universalists and the relativists are really talking about different things. When the universalists talk about emotion, many have uppermost in mind the physiological response (the cascade of hormones and neurotransmitters, the shifts in facial expression and blood pressure), while the relativists have uppermost in mind the cultural category labeled by an emotion term (e.g., “Is Balinese lek the same as American English shame?”).
In any event, the particular theoretical posture of the investigators of emotion does not prevent them from teaching us something interesting about the human experience and expression of emotions. Thus Lutz (1987, pp. 294–297) can tell us how emotional displays can initiate culturally scripted exchanges of emotion among the Ifaluk (e.g., *ker* “happiness/excitement” begets *song* “justifiable anger” begets *metagu* “fear/anxiety”); Gerber (1985, pp. 155–158) can explain how and why both Mead and Freeman got it right and wrong on the Samoans’ alternately suppressed and expressed rage at authority; and Ekman (2003) can articulate which facial muscles are attended to when people distinguish fake smiles from real ones. All are interesting observations despite the theoretical differences among the observers.

2. Methods of assessing cultural emotion systems

Rather than argue about whether emotions are universal or culturally specific, it is more productive to make use of these observations and explore the nature of the similarities and differences among cultural emotion systems. In order to do this, one must first address the methodological question. What does one learn from different ways of assessing the similarities and differences in cultural emotion systems? One can discern two distinct methods, which I will call the method of translation and the method of mapping.\(^6\)

2.1. The Method of Translation

In the method of translation, the focus of the analysis is on the semantic relationships among a set of emotion terms. One proceeds by first finding equivalent sets of emotion terms in two (or more) languages. One then asks native speakers to judge the similarity

\(^6\) There is a third method employed by many ethnographers [e.g., Briggs (1970), Levy (1973), Rosaldo (1980), Lutz (1988)] which might be called “the method of empathy.” The Oxford English Dictionary defines empathy as “the power of projecting one’s personality into (and so fully comprehending) the object of contemplation.” This aptly describes what many ethnographers attempt to do to enter into an alternative cultural emotional world, if “personality” is generously interpreted to include the cultural and linguistic systems of the empathizer. Lutz (1985, pp. 68–69) explains:

The claim has been made that no ethnopsychological system is ever explained “in its own terms”; to say that ethnotheory can be explored without reference to other theoretical systems is to claim that the anthropologist goes to the field as a tabula rasa, that is, without an ethnopsychological interpretive system of her or his own. In the translation of ethnopsychologies, we rely heavily on our own and other’s understandings of concepts such as ‘mind’, ‘self’, and ‘anger’. The invisibility of culture (the taking of belief for knowledge or of language for the object) has left opaque the extent to which attributions of ‘anger’ are cultural attributions. This captures a necessary and probably inevitable aspect of ethnographic fieldwork – one can no more approach and successfully understand an unfamiliar cultural system as a blank slate than one could learn a second language without being influenced by the phonological, semantic, and grammatical categories of one’s first language. Indeed, the “confusionism” described by Shweder and Bourne (1984) occurs when empathy and comprehension fail [cf. Sperber (1982)]. However, as a method of comparing cultural emotion systems, the method of empathy has much in common with the method of translation – one starts by finding the categories in the other system that best match those that are salient in one’s own. Of course, in the end, the ethnographer is likely to be struck by the categories and responses that are the most different from her own and forget about those that are similar.
of the meanings of the terms. Finally, one assesses the degree to which the semantic relationships among the terms are similar for the two languages [e.g., Russell (1983, 1989), Heider (1991), Romney et al. (1997)]. In the translation method, translation is a prerequisite to the rest of the analysis.

The translation method was used by Russell (1983, 1989) to argue that the emotion space has a circumplex structure. Figure 1 shows the circumplex structure of American English emotion terms, as judged by American English speakers [Russell (1989)] and a translation equivalent set of Japanese emotion terms, as judged by Japanese speakers [Russell (1983)]. The vertical (arousal) dimension ranges from low-arousal terms at the bottom of the figure (e.g., tired, sleepy, drowsy, calm) to high-arousal terms at the top (e.g., alarmed, tense, angry, aroused, astonished, excited). The horizontal (evaluation) dimension ranges from unpleasant emotions at the left (e.g., miserable, distressed, gloomy, sad, frustrated) to pleasant emotions at the right (e.g., delighted, happy, pleased, glad). The vectors in Figure 1 show the difference between the American English and the Japanese judgments: the head of the vector is the placement in the circumplex of the American English term and the point of the vector is the placement of the Japanese term. The circumplex of Japanese terms is rotated an average of 23° counterclockwise relative to the circumplex of English terms. Russell (1983) also elicited similar circumplex structures from the judgments of Chinese, Croatian, and Gujarati speakers.

Fig. 1. Circumplex structures of American English and Japanese emotion terms. Vectors show displacement of Japanese terms (pinpoints) relative to the English terms (pinheads). Data are from Russell (1983, 1989).
The method of translation is also useful to compare patterns of synonymy within and between languages, as demonstrated by Heider (1991). Heider presented multilingual speakers of three Sumatran languages (Bahasa Indonesia, Bahasa Minangkabau, and Javanese) with a set of target emotion words and asked them to list synonyms in each language and translation equivalents in the other languages they knew. The fruits of his method are illustrated in Figure 2. His method very clearly demonstrated that emotion words in one language are not neatly equivalent to emotion words in another language. So, for example, the term *curiga* (‘suspicious’) in Bahasa Indonesia corresponds to cognate terms *curiga* and *curigo* in Bahasa Minangkabau and Javanese, respectively. But two other terms in Bahasa Minangkabau are regarded as translation equivalents of the Bahasa Indonesia term *curiga* (*sak ati, badatak ati*), while two additional terms in Bahasa Indonesia are regarded as translation equivalents of the Bahasa Minangkabau term *curiga* (*syak wasangko, prasangka*). Of these, the former, but not the latter, is regarded as the equivalent of the Javanese term *curigo*. Clearly, there is not a one-to-one mapping of the meanings of emotion terms in the three languages, even if the terms are cognate.

---

Fig. 2. Between-language translation equivalents of indecision emotion terms in Bahasa Minangkabau (top), Bahasa Indonesia (middle), and Javanese (bottom). From Heider (1991, Figure 16.4). Copyright 1991 by Cambridge University Press. Reprinted with permission.

---

7 There is an interesting pattern in Heider’s (1991) data, in which one category in a cluster appears to serve as the prototype of the cluster. For example, the terms *bimbang, sangsi,* and *ragu* look as though they are exact cognates of each other in Bahasa Minangkabau and Bahasa Indonesia. However, speakers of both languages pick the term *ragu* as the translation equivalent of all three terms in the other language. *Curiga* and *cameh/cemas* also appear to be prototypes of their respective clusters. Evidently, these bilingual speakers maintain different mental maps of the meaning relations in these two languages, even though the languages are quite closely related. In other words, the semantic systems of the two languages do not seem to have converged in the minds of their bilingual speakers.
However, the translation method is not adequate to show that speakers of different languages “share a single model of the semantic structure of emotion terms,” as claimed by Romney, Moore and Rusch (1997, p. 5489) in their comparison of the judgments of similarity of English and Japanese emotion terms. They chose sets of emotion terms in English and Japanese in such a way as not to privilege either language over the other, found the best translational equivalent of each term in the other language, and thus arrived at sets of terms that were judged as translational equivalents in the two languages. They then asked native speakers of English and Japanese to judge the similarity of the terms. The English and Japanese judged the similarity of the meanings of the terms similarly, as shown in Figure 3.

The 97.5 confidence ellipses surrounding the mean position of the terms for the two languages are impressively small; English and Japanese speakers appear to perceive the same pattern of similarity among the terms. The dimensions of the space appear to be similar to those found by Russell (1983), with evaluation as the horizontal axis and arousal as the vertical axis. The fact that the terms in Romney et al.’s (1997) analysis do not arrange themselves in a circumplex, whereas they do in Russell’s (1983) analysis, is probably due to the fact that Russell’s analysis normalizes the data and Romney et al.’s does
not. However, I would argue that this finding tells us relatively little about whether or not Japanese and English have the same “semantic structure” underlying their emotion terms.

One problem is that the “semantics” of emotion terms revealed by a similarity judgment task is relatively shallow – one can complete the task using just the two dimensions of Russell’s circumplex: evaluation and arousal. Osgood found similar dimensions (evaluation, potency, and activity) structuring informants’ responses to his “semantic differential” [Osgood, Suci and Tannenbaum (1957), Osgood, May and Miron (1975)], but I do not believe that either he or anyone reading him would claim that he had shown that the diverse domains all shared the same semantic structure. Instead, what Osgood and his collaborators showed is that these three dimensions of connotative or affective meaning have broad relevance across a number of domains, regardless of the substantive nature of the domain. In other words, these dimensions do not tell us much about emotions in particular, since a great many domains can be placed in this connotative space. Shweder and Bourne (1984, p. 160) have made this point bluntly but aptly: arguing that at a sufficiently general level of discourse, “‘God’ and ‘Ice Cream’ are descriptively equivalent; both are perceived as good, strong, and active.” Consider three emotional states: being in a paroxysm of grief, being paralyzed by terror, and being in a towering rage. All three are very distinct highly aroused negative emotions that would fit into the same spot on the circumplex. Obviously, the two dimensions of evaluation and arousal are not sufficient to represent important semantic distinctions in this set of terms; many more semantic features are required to distinguish between intense grief, terror, and rage.

Although Romney et al. did not find a circumplex structure in their analysis, the same problem is evident in an examination of Figure 3: disgust, anger, and hate all overlap, as do tired and bored.

However, the deeper problem here is one pointed out many years ago by Eric Lenneberg (1953) in his critique of the methods used by Benjamin Lee Whorf to argue that speakers of different languages are coerced into radically different understandings of the world by the nature of their native languages. Figure 4 illustrates Whorf’s method. Whorf would take a sentence in English (e.g., He invites people to a feast.), find its translation in another language, translate the morphemes of the translation back into English, and then boggle at the absurdity (in English) of the resulting string of translated morphemes (e.g., ‘Boil – ed – eat – ers – go-for – he does’). The difficulty with this method, observed Lenneberg, is that if the other language represents a wildly different way of understanding the world, one loses that worldview as soon as one has rendered it into English. The reason that this problem is fatal to the claims of Romney et al., but does not seriously harm the claims of Russell and Heider, has to do with the nature of the claims made, and the role of translation in their procedure.

For Heider, the translation is itself the data analyzed, and so has a status like that of similarity judgments made in any one language; they are judgments of the degree to which a pair of terms in the same or different languages are similar in meaning. Heider also uses the analysis of the data to advance a relativist interpretation; he mainly notes

---

8 Shweder (1994, pp. 43–44) has a thoughtful discussion of a similar observation.
the differences in the relations of the patterns of synonymy of terms that are regarded as translational equivalents. (It is difficult to see how the data he collects could be used to do anything but advance a claim of cultural difference.) So his claims do not depend on the notion that the terms are exact equivalents of each other. In fact, he claims that the terms are not equivalent, if one looks at their relations of synonymy and translation equivalence to other terms.

The threat to the validity of Russell’s claims is somewhat greater, because he uses the translation as a preliminary to similarity judgments of the terms; but his principle claim is that the emotion terms form a circumplex with two dimensions: arousal and evaluation. This claim would hold true even if the terms in the languages compared were not translation equivalents. (The more serious threat to the validity of Russell’s claims is the suggestion that the circumplex structure of the emotion space he discovers is an artifact of normalizing the data matrices.)

However, for Romney et al. (1997), the translation is preliminary to the subsequent data collection and analysis, which depend on the exact equivalence of the terms translated. First, informants judge the similarities of the meanings of the terms, next the similarity structures are compared and found to be in concordance, and finally it is concluded that speakers of different languages “share a single model of the semantic structure of emotion terms” (1997, p. 5489). However, the game has been given away at the first step, because the procedure presumes the answer to the research question...
itself. The degree of correspondence between emotion systems cannot be assessed if one presumes to know the corresponding terms from the outset – one simply “discovers” one’s assumptions. Translation should be a result of the analysis, not its first step.

One way of illustrating this problem is to compare the structures in the representations of the Japanese and American English similarity judgments collected by Russell (1983), shown in Figure 1, with the representation of those same judgments, as collected by Romney et al. (1997), shown in Figure 3. Seven terms (and their Japanese equivalents) are used by both Russell and Romney et al.: anger, excitement, fear, tired, bored, sad, and happy. Figure 5 shows the difference between Russell’s (1983, 1989) and Romney et al.’s (1997) graphic representations of these seven American English and corresponding Japanese emotion terms. If one takes the coordinates of the seven emotion terms from the figures, computes the Euclidean distances among them, and compares the three structures, one discovers that English (as assessed by Romney et al.)

---


---

9 To be clear, my critique of Romney et al. (1997) has to do with the methods employed in the data collection and analysis and not with their substantive results. Their conclusions may very well be valid even if their methods are unreliable.
is more different from English (as assessed by Russell) than it is from Japanese (as assessed by Romney et al.); and that Japanese (as assessed by Romney et al.) is more different from Japanese (as assessed by Russell) than it is from English (as assessed by Romney et al.). Table 1 shows this comparison. The discrepancy is probably due to the difference in methods noted above. Nevertheless, this example illustrates why one should exercise caution in interpreting these figures as representing the semantic structures of the languages.

Table 1

<table>
<thead>
<tr>
<th></th>
<th>r</th>
<th>QAP Z</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Romney et al. to Russell English</td>
<td>0.56</td>
<td>2.7</td>
<td>0.0057</td>
</tr>
<tr>
<td>Romney et al. to Russell Japanese</td>
<td>0.64</td>
<td>2.9</td>
<td>0.0031</td>
</tr>
<tr>
<td>Russell English to Russell Japanese</td>
<td>0.87</td>
<td>3.8</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

The second general method of assessing the similarities and differences of cultural emotion systems is the method of mapping. In this method, the focus of the analysis is on the mapping of emotion terms to a collection of referents. One proceeds by presenting the same emotionally evocative stimuli to native speakers of (at least) a pair of languages. One asks the native speakers to identify the emotion expressed or evoked in the emotional stimuli. Finally, one assesses the degree to which the implied similarity of the stimuli is similar for the two languages. In the mapping method, translation is a consequence of the analysis.

The method of mapping is the research tactic predominantly used in the cross-cultural comparison of color classification [e.g., Berlin and Kay (1969), Kay et al. (1997)] and in ethnobiology [e.g., Boster et al. (1986), Boster and D’Andrade (1989)]. One important difference between those domains and that of emotions is that actual items from the domain can be presented to an informant to be identified. That is, one can show informants a color or a bird, a fish or a mammal, and ask them what it is. Emotions do not have
an independent existence in the same way – one cannot reach in and wrench an emotion out of a human heart in order to present it as a disembodied thing free of the individual experiencing the emotion. The best one can do is to proceed by indirection, picking stimuli that have clear emotional valences. One place to start is with facial expressions or gestures\textsuperscript{11} of emotion. Facial expressions have the advantage of being easily presented to informants as stimuli and there is a considerable literature suggesting that facial expressions are of universal human emotional importance, even if there is cultural variation in how they are interpreted. It is this cultural variation that one wants to discover.

An example of the use of this method is the facial expression naming task. In this task, 22 photographs of an actor and an actress expressing emotions selected to evenly sample Russell’s circumplex emotion space are presented to native speakers of a variety of languages. Figures 6 and 7 show the photographs used in this task. The informants are asked to give a word or short phrase that best describes how the man/woman was feeling at the time the photograph was taken. The responses were recorded. It was necessary to clean up the data by collapsing various forms of the same root (e.g., anger = angry; sadness = sad, etc.) and weeding out rare or idiosyncratic responses. Collapsing did not extend to judgments of synonymy, so mad and angry were left as separate responses. One can then build a mapping matrix that represents the mapping of terms (rows) to the photographs (columns). Table 2 shows a portion of the American English mapping matrix, showing the mapping of a subset of terms to the male faces. The columns of the mapping matrix are correlated to infer the similarity of the faces implied by the mapping. Finally, one can compare the correlation matrices to see how similar the mappings of terms are across the various languages.

This approach differs from the Facial Expression Program [e.g., Tomkins (1962), Izard (1971), Ekman et al. (1972)] in two important ways. First, informants can freely choose whatever description they like of the emotion expressed by the actor in the photograph rather than being forced to choose among a limited and predetermined set of options. Second, the analysis focuses on the mapping of a whole field of terms onto the range of facial expressions rather than on whether a particular term refers to its “correct” facial expression (e.g., “Is the term ‘fear’ applied to the ‘fear-face’?”).

A comparison of the inferred similarities of the photographs in the various languages supports a middle-of-the-road position. There is evidence of both strong commonalities and important cultural differences in the ways in which the terms in the various languages map to the facial expressions of emotion (average Pearson $r = 0.74; p < 0.00001$), with roughly half the variance shared and half unique to each language. Figure 8 shows an

\textsuperscript{11} Some would call the stimuli in the facial expression naming task photographs of “facial gestures” rather than of “facial expressions,” distinguishing between those movements of facial muscles that express a genuine emotion by the individual (a facial expression) from an actor’s deliberate posing of the facial muscles to simulate an emotional expression (a facial gesture). Because the actors in the photographs are merely posing rage, fear, boredom, etc., it cannot be said that they are expressing emotions, only that they are posing them. However, this distinction is not generally adhered to. Some assign a different meaning to the term “facial gestures,” others treat “facial gestures” and “facial expressions” as exact synonyms. I use the term “facial expressions” because it is more familiar to a majority of readers.
average-link clustering of the languages based on the correlations among the mappings shown in Table 3. It would seem to indicate an effect of language group, with all of the Indo-European languages mapping to the faces more similarly to each other than to the one non-Indo-European language, Shuar (or Jívaro).

However, the apparent clustering may also be a result of sample size, such that languages with larger sample sizes of informants are more robustly estimated and agree with each other more than those with smaller samples of informants. Figure 9 shows the nearly linear relationship between the log sample size of the languages with a measure of the average agreement of each language to the others\(^\text{13}\) \((r = 0.91; p < 0.05)\). If the

---

\(^{12}\) These photographs (and those shown in Figure 7) are a subset of those originally used in a study of emotion recognition by left and right hemisphere stroke patients in collaboration with Lee Xenakis Blonder. Actors were asked to pose the states of being astonished, delighted, happy, satisfied, relaxed, bored, sad, annoyed, angry, and afraid.

\(^{13}\) The measure of agreement is the first factor score from a factor analysis of the interlanguage correlation matrix, which is roughly equal to the square root of the average correlation of that language with the other languages.
pattern of language clustering is an artifact of sample size, larger samples would lead to even higher correlations among the mappings. An answer to this question would require that the facial expression naming task be administered to large numbers of speakers of several non-Indo-European languages.

The next step is to stack the mapping matrices one on top of another and to do a correspondence analysis of the stacked matrices to see how terms in the different languages map to the faces [Weller and Romney (1990, pp. 85–91)]. Again, the columns of this overall stacked matrix correspond to the photographs of the emotion faces and the rows to the terms in each of the languages compared. Figure 10 shows a plot of the facial expressions on the first two dimensions of the correspondence analysis, with capital letters to refer to the series of 11 male facial expressions and lowercase letters to refer to the photographs of the 11 female facial expressions. To orient the reader, “C” and “K” are the male facial expressions at 11 and 12 o’clock, respectively in Figure 6, while “g” and “b” are the female facial expressions at 8 and 9 o’clock, respectively in Figure 7. Figures 11, 12, 13, 14, and 15 show how the emotion terms in English, Spanish, Polish, Italian, and Shuar, respectively, map to the faces shown in Figure 10. As one can see by examining Figure 11, the
mapping of the terms roughly approximates Russell’s circumplex structure, with fright terms at the top (frightened, scared, fear) moving clockwise into terms referring to pure arousal (shocked, surprised), thence to aroused positive affects (ecstatic, delighted, elated), to arousal-neutral positive affects (pleasant, happy, amused), to low-arousal positive affects (content, satisfied), to low-arousal neutral affects (normal, nothing, pensive),
to low-arousal negative affects (tired, bored), to aroused negative affects at the top left of Figure 11 (sad, worried, disgust, upset). The one major anomaly is that some negative aroused emotions (angry, mad) fall together with the low arousal negative affects in the lower left quadrant of Figure 11. Another is that the whole structure is rotated about 30° clockwise from the canonical orientation of Russell’s circumplex. The same general circumplex structure and anomalies can be found in the other languages as well.

One can make use of the correspondence analysis of the stacked mapping matrix to find the terms in other languages that most closely approximate the coordinates of a term in a target language. In this way, I can pay off on the promise that translation will be the result

<table>
<thead>
<tr>
<th>Language (N)</th>
<th>English</th>
<th>Spanish</th>
<th>Polish</th>
<th>Italian</th>
<th>Shuar</th>
</tr>
</thead>
<tbody>
<tr>
<td>English (118)</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish (61)</td>
<td>0.88</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish (33)</td>
<td>0.77</td>
<td>0.77</td>
<td>1.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian (29)</td>
<td>0.77</td>
<td>0.75</td>
<td>0.75</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Shuar (19)</td>
<td>0.67</td>
<td>0.71</td>
<td>0.56</td>
<td>0.53</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Fig. 9. Scatter plot of languages in facial expression naming task showing increase in average correlation with other languages with increasing log sample size ($r = 0.91, p < 0.05$).
Fig. 10. Plot of male (in caps) and female (lower case) facial expressions using column coordinates from correspondence analysis of crosslanguage mapping matrix.

Fig. 11. Plot of English emotion descriptors using English row coordinates from correspondence analysis of crosslanguage mapping matrix.
Fig. 12. Plot of Spanish emotion descriptors using Spanish row coordinates from correspondence analysis of crosslanguage mapping matrix.

Fig. 13. Plot of Polish emotion descriptors using Polish row coordinates from correspondence analysis of crosslanguage mapping matrix.
Fig. 14. Plot of Italian emotion descriptors using Italian row coordinates from correspondence analysis of crosslanguage mapping matrix.

Fig. 15. Plot of Shuar emotion descriptors using Shuar row coordinates from correspondence analysis of crosslanguage mapping matrix.
of the analysis rather than its premise. However, this is a translation by virtue of shared reference rather than shared meaning; nothing guarantees that the terms that are the closest to each other with respect to how they refer to the faces will have exactly the same meaning or be the terms that dictionaries offer as translation equivalents of each other. Table 4 shows the terms in Spanish, Polish, Italian, and Shuar that have coordinates that are close to those of the English basic emotion terms *angry*, *fear*, *happy*, *sad*, and *surprised*. These are the terms that are applied to the photographs most similarly to the way that the corresponding English terms are applied. Table 5 shows the translations of the same terms as given by dictionaries. In Table 4, the terms that are those also given by a dictionary are shown in **bold**.

Note that for the Indo-European languages, this way of choosing translation equivalents results in the dictionary definition three-fifths of the time in all three languages; not bad, but not perfect. However, the method fails to deliver the dictionary definitions in Shuar for all but one term, *warawai* (‘happy’). A closer examination of this case yields interesting insights into how the Shuar approached the facial expression naming task.

A very large proportion of the American English speakers’ responses to the facial expressions were emotion terms, that is, terms which refer primarily to internal subjective states such as anger, sadness, fear, etc. While the Shuar sometimes responded to the facial expressions with terms that can also be regarded as emotion terms proper, these made up only about half of their total responses. The other responses fell into a number of broad classes: (1) flat descriptions of the facial expression or emotional behavior itself (‘She’s smiling,’ ‘He has his mouth open,’ ‘She has gritted her teeth’); (2) descriptions of other actions (‘He’s saying “Who is there?”’, ‘He’s making a joke’); (3) descriptions of beliefs or thoughts with an emotional valence (‘He’s thinking of his lover,’ ‘He’s thinking he will never find a wife’); (4) descriptions of events with an emotional valence that have happened to the actor (‘Her mother just died’). Of these classes, the flat description of the facial expression or emotional behavior itself was the most general and provided a nearly complete partition of the facial expression space, as shown in Figure 16. Note also that three of these terms *etser*, *uutiawai*, and *wanka* are the ones that have the closest coordinates to the English terms *angry*, *sad*, and *surprised*, respectively, as shown in Table 4. The Shuar informants were not unique in offering descriptions of emotional behaviors as responses to the photographs; *usmiejch*, the Polish term closest in reference to *happy*, refers to smiling while *piangere*, the Italian term closest in reference to *sad*, refers to weeping.

So what is going on here? One possibility is that the Shuar have a folk psychology similar to what Lutz (1987) has described for the Ifaluk. According to Lutz, the Ifaluk locate emotions as occurring primarily among people rather than within them. Similarly, it appears that the Shuar often interpret facial expressions in terms of publicly observable interactions and behaviors rather than in terms of internal psychic states, as Americans

---

14 If one wants translation by shared reference to more precisely match the semantics of the terms in the languages translated, the stimuli named have to capture the semantic contrasts, something that photographs of the facial expression of emotion cannot do. One would need to use descriptions of emotional scenarios that more accurately reflect the semantic distinctions made by the terms in order for the technique to produce something closer to translation equivalents.
Table 4
Translation by shared reference in the facial expression naming task

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>Polish</th>
<th>Shuar</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>coraje</td>
<td>z ość</td>
<td>etser</td>
<td>aggressiva</td>
</tr>
<tr>
<td>fear</td>
<td>impresionada</td>
<td>strach</td>
<td>kakaratamia</td>
<td>paura</td>
</tr>
<tr>
<td>happy</td>
<td>feliz</td>
<td>uśmiech</td>
<td>uutiaiawai</td>
<td>contento</td>
</tr>
<tr>
<td>sad</td>
<td>triste</td>
<td>żal</td>
<td>uutiaiawai</td>
<td>piangere</td>
</tr>
<tr>
<td>surprised</td>
<td>sorpresa</td>
<td>zdziwienie</td>
<td>wanka</td>
<td>sorpresa</td>
</tr>
</tbody>
</table>

Table 5
Translation by dictionary

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
<th>Polish</th>
<th>Shuar</th>
<th>Italian</th>
</tr>
</thead>
<tbody>
<tr>
<td>angry</td>
<td>enojo</td>
<td>z ość</td>
<td>kajeawai</td>
<td>rabbia</td>
</tr>
<tr>
<td>fear</td>
<td>miedo</td>
<td>strach</td>
<td>sapijmiawai</td>
<td>paura</td>
</tr>
<tr>
<td>happy</td>
<td>feliz</td>
<td>radość</td>
<td>warawai</td>
<td>contento</td>
</tr>
<tr>
<td>sad</td>
<td>triste</td>
<td>smutny</td>
<td>kuntuts</td>
<td>triste</td>
</tr>
<tr>
<td>surprised</td>
<td>sorpresa</td>
<td>zdziwienie</td>
<td>awaknamkai</td>
<td>sorpresa</td>
</tr>
</tbody>
</table>

![Shuar](image.png)

Fig. 16. Plot of Shuar descriptors of facial behaviors.
generally do. Like the Ifaluk, the Shuar folk model of the mind does not distinguish between thoughts and feelings; both are anentaimia. Perhaps the absence of a lexical distinction between thoughts and feelings gives descriptions of the actor’s thoughts and beliefs an equal standing with descriptions of internal emotion states as possible accounts of what the actor was thinking/feeling at the time the photograph was taken.  

Although the possibility of a hint of Micronesia in the rainforests of Amazonia is tantalizing, there are alternative explanations. The photographs of the actors’ facial expressions are disembodied and liberated from any defining context. The decontextualized disembodiment of the faces gives the informants three options in a response: (1) to guess the general emotion without trying to figure out the precipitating event; (2) to retreat and not pick an emotion term but instead offer a flat description of the facial expression itself; or (3) to invent a social context that would explain how the actor came to feel a particular way and display that particular expression. In this interpretation, the reason why the Shuar respond differently from the Indo-Europeans and choose options 2 and 3 as often as option 1 has less to do with the details of their folk psychology and more with their experience with schooling. The Shuar have much less exposure to decontextualized problem solving in which the student has to answer a question based solely on a set of presuppositions supplied by the teacher, while the mainly urban, university-educated informants in the American English, Mexican Spanish, Polish, and Italian samples were familiar with this kind of naming task, because they had frequently encountered similar ones at school. The Indo-European informants were usually able to supply the name of an emotional state (option 1) because that is what the task calls for, even if the photograph itself does not supply enough context to really disambiguate what the actor is feeling. A scowl may be the product of envy, jealousy, being slighted, or a variety of other causes; one is unable to tell which from the picture alone. That is why the Indo-European informants usually supply generic “basic” emotion terms which do not require them to choose among these alternatives – they say the actor is angry, but do not guess what he might be angry about. In other words, the Shuar can be interpreted as displaying an “empirical bias” of the sort that Scribner (1977) found in her review of many studies of reasoning among unschooled informants in traditional societies. For example, Scribner (1977, p. 491) asked an adult Vai (Liberian) informant to solve the following syllogistic reasoning problem: “All people who own houses pay house tax. Boima does not pay a house tax. Does he own a house?” Her informant freely invented new information in offering an answer: “Boima has a house but he is exempted from paying house tax. The

---

15 The prompt used to ask the Shuar what the actor in the stimulus photographs is feeling was Wari anentaimiawai?, literally “What is he thinking/feeling?” [Note: should the word be anentaimiawai?]  

16 Fernández-Dols and Carroll (1997) review a substantial literature showing that facial expressions are dependent on their context for their full interpretation and that the same expression can be interpreted differently in different contexts. They state (1997, p. 289):  

Those investigations that supported the assumption of the self-sufficiency of facial expressions were based on an experimental paradigm that misrepresented the contextual sources of information. The recognition of emotions from face [faces?] is, as is any other perceptual process, no exception to the principles relating figure to ground.
government appointed Boima to collect house tax so they exempted him from paying house tax.” The Shuar informants are similarly importing new information, presumably to make up for a perceived poverty of information and context in the stimulus itself. Further research would be required to be confident in one interpretation or another.

This problem of missing meaning is a problem for informants and ethnographers alike; both want to recover the propositional content of emotion terms. I think some of the frustration that the cultural relativists feel about the results obtained by the universalists is that, by focusing on the facial expression of emotion and the “basic” emotions associated with them, the universalists miss out on much of the meaning or propositional content of emotion lexicons and hence much of what emotion terms are all about. Emotions have clear provocations and sequelae, they are important signals both intra- and interpersonally. What is the best way to explore this missing meaning?

3. Theories of emotion

I think that to develop the right sort of method, it helps to have a definition and a theory of emotion that makes these provocations and sequelae part of what one explores when investigating a cultural emotion system. I think the appropriate theory and definition come from the intersection of appraisal theories [e.g., Smith and Ellsworth (1985), Lazarus (1991)] and evolutionary theories of emotion [e.g., Plutchik (1980), Frank (1988), Nesse (1990), Cosmides and Tooby (2000)]

17. Bridging these two approaches, Johnstone and Scherer (2000, 220) define an emotion as “a phylogenetically evolved, adaptive mechanism that facilitates an organism’s attempt to cope with important events affecting its well-being.” One of the attractive features of this definition is its very vagueness; it can be interpreted to include both the aspects of emotion which may be universal and part of the evolved human biological endowment (e.g., physiology and qualia) and the aspects which are locally culturally constructed (e.g., the cognitive appraisals and the linguistic categories of emotions). More important, it emphasizes that, in general, emotions function to serve the interests of the organism experiencing and expressing them. Both evolutionary and appraisal theorists agree that emotions are about something; they help the organism to respond appropriately and adaptively to problems and opportunities posed by events in the world. The propositional content of emotion terms in various emotion lexicons is often about the nature of those problems and opportunities.

The compatibility of the two approaches can also be seen in the overall architecture of the flow between event and action that proponents of evolutionary and appraisal theory envision. For example, Table 6 shows Plutchik’s outline of the key elements in the emotion sequence [modified from Plutchik (2000, p. 69)], while Table 7 shows the core relational themes identified by Lazarus (1991, p. 122) for the same emotions. Both

17 The unification of psychodynamic and evolutionary approaches in psychology is an old ambition, dating back at least to Bowlby (1969).
18 For a nice compendium of alternate definitions of emotion, see Plutchik (2003, p. 18).
attribute the same functions to this suite of emotions. The sequence outlined by Plutchik (stimulus – interpretation – affect – behavior) also corresponds to the “commonsense” sequence linking events and actions identified by Ellsworth (1991) and to that described in D’Andrade’s (1987) account of the American folk model of the mind, illustrated in Figure 17.

Armed with an approach to emotion that can encompass both the evolved universal biological endowment and the locally culturally specific understandings of emotional experience, we next need a method to discover the propositional content of emotion terms that will allow us to properly compare different cultural emotion systems. Anna

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>Cognition</th>
<th>Feeling State</th>
<th>Overt Behavior</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threat</td>
<td>“Danger”</td>
<td>Fear</td>
<td>Escape</td>
<td>Safety</td>
</tr>
<tr>
<td>Obstacle</td>
<td>“Enemy”</td>
<td>Anger</td>
<td>Attack</td>
<td>Destroy</td>
</tr>
<tr>
<td>Loss</td>
<td>“Abandonment”</td>
<td>Sadness</td>
<td>Cry</td>
<td>Reattach</td>
</tr>
<tr>
<td>Gain</td>
<td>“Possess”</td>
<td>Joy</td>
<td>Retain</td>
<td>Gain resources</td>
</tr>
</tbody>
</table>

Table 7
Core relational themes of emotions according to Lazarus. Adapted from Table 3.4 in (Lazarus, 1991, p. 122)

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Core Relational Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fright</td>
<td>Facing an immediate, concrete, and overwhelming physical danger.</td>
</tr>
<tr>
<td>Anger</td>
<td>A demeaning offense against me and mine.</td>
</tr>
<tr>
<td>Sadness</td>
<td>Having experienced an irrevocable loss.</td>
</tr>
<tr>
<td>Happiness</td>
<td>Making reasonable progress toward the realization of a goal.</td>
</tr>
</tbody>
</table>

Fig. 17. D’Andrade’s representation of the American folk model of the mind. From D’Andrade, p. 162. Copyright 1995 by Cambridge University Press. Reprinted with permission.
Wierzbicka (1992) describes one approach to this problem in pursuing a goal very similar to ours, to discover “universal human concepts in culture-specific configurations.” Her way of comparing the propositional contents of emotion terms in different languages is to first translate them into a set of semantic primitives that comprise her universal semantic metalanguage.

According to Wierzbicka, the semantic primitives should be clear and self-explanatory; distinct and indefinable; used in constructing more complex concepts; useful in many languages; and lexical universals [Wierzbicka (1992, pp. 11–12)]. She illustrates the approach by rendering the meanings of an Ngiyambaa (Austronesian) avoidance term kuyan and the English term ashamed in her metalanguage as follows [Wierzbicka (1992, pp. 132–133)]:

**kuyan**
- X thinks something like this
- I am near person Y
- this is bad
- something bad could happen because of this
- people could think something bad about me because of this
- I don’t want this
- because of this, X feels something
- because of this, X wants to do something
- X wants not to be near this person

**(X is) ashamed**
- X thinks something like this
- people can know something bad about me
- because of this, people can think something bad about me
- I don’t want this
- because of this, I would want to do something
- I don’t know what I can do
- because of this, X feels something bad

The attraction of Wierzbicka’s method is that it makes it easy to see the similarities and differences between the experiences of kuyan and ashamed; it is also plausible that one could find the semantic primitives that she has employed in many, if not most, world languages. The drawback is that the method depends on the analyst’s own understanding of things – it is Wierzbicka, herself, who renders all of her examples in her semantic metalanguage. Although her rendering of ashamed looks plausible enough, I am doubtful that I would have come up with exactly the same rendering even though I am a native English speaker. This is perfectly adequate if the goal is simply to compile a dictionary, but Wierzbicka wants the difference between her definitions of terms in different languages to capture the difference in how the speakers of those

---

19 The proposed set of semantic primitives would not be expected to be found in all human languages, however. Many of them are missing from Wao tededo (the language of the Waorani), for example.
languages understand the world. What is sufficient for lexicography is insufficient for ethnology. Even if she exactly captures the difference in the central tendencies of the meanings of the terms in the two languages, we still need some measure of the variation in the understanding of those terms in the two language communities to assess the significance of the difference. In other words, we need the same information as in a statistical test, to see whether the difference between the systems is large in comparison to the variation within them. Without this additional information, we risk reifying non-Western emotion categories, along with Western ones [cf. Lutz (1985), pp. 38–39, Wierzbicka (1992). pp. 119–121]20.

4. Cross-cultural scenarios as a tool to compare emotion categories

My solution to this problem is to try to unpack the meanings of emotion terms much as Wierzbicka attempts to, not by depending on my own insights but instead by allowing a sample of native speakers decide what the terms mean. This method allows us to get both a sense of the central tendency of the meanings of emotion terms and a measure of the variation in the understandings of those meanings in a way that is more independent of the ethnographer’s judgments. A pilot project exploring this method was carried out with the Waorani of Eastern Ecuador. The initial step in this process was to use the facial expression naming task to determine the most frequently used emotional descriptors of the facial expressions and for each descriptor to ask “what causes you to feel _____?”21 For example, the male facial expression “C” at 11 o’clock of Figure 6, which American English speakers described as frightened, scared, or fear, was described as ankai giñente by my

20 A consequence of Wierzbicka’s rendering of emotion terms into her semantic metalanguage is that the definitions become quite wordy, because she must avoid the use of more complex terms that concatenate several semantic features into a single lexical item. For example, she renders schadenfreude as follows (1999, pp. 104):

(a) X felt something because X thought something
(b) sometimes a person thinks about someone else:
  (c) “many good things happened to this person before now
  (d) this person thought: ‘this is good’
  (e) something bad has happened to this person now
  (f) now I think: this is good”
  (g) when this person thinks this, this person feels something good
  (b) X felt something like this
  (i) because X thought something like this

The definition is so prolix and confusing (whatever its other virtues) that it is difficult to imagine presenting it to native German speakers to have them evaluate whether she got it right. On the other hand, defining schadenfreude as “feeling good when something bad happens to someone who has been fortunate until now” is sufficiently terse to hold in memory and judge its adequacy as a definition.

21 I do not advocate using facial expressions of emotion to generate lists of emotion terms in general, as one is likely to generate only terms for the most general, basic affects and not generate any of the most specific emotion terms. For example, it is much more likely that a facial expression will be described as angry rather than jealous or envious. However, the method proved a useful and adequate way to proceed with Wao tededo, which has a very restricted psychological and emotional lexicon.
Waorani informants. When Waorani informants were asked “what causes you to feel ankai giñente?” they answered with events such as “seeing a jaguar in the forest” or “seeing a poisonous snake on the path.” The causes of ankai giñente as judged by the Waorani fit fairly well with the core relational theme that Lazarus (1991, p. 122) attributes to fright: “facing an immediate, concrete, and overwhelming physical danger.” Similarly, the female facial expression “a” at 10 o’clock of Figure 7, which American English speakers described as angry or mad, was described as piinte by my Waorani informants. When Waorani informants were asked ‘what causes you to feel piinte?’ they answered with events such as ‘someone stealing my things,’ ‘someone passing close to my house without greeting me,’ and ‘coming home and my wife does not give me manioc beer.’ The causes of piinte as judged by the Waorani also fit fairly well with the core-relational theme that Lazarus (1991, p. 122) attributes to anger: “a demeaning offense against me and mine.” As a final example, the male facial expression “J” at five o’clock of Figure 6, which American English speakers described as content, pleased, or satisfied, was described as gane ponente (literally ‘cool thinking’) by my Waorani informants. When Waorani informants were asked ‘what causes you to feel gane ponente?’ they answered with events such as ‘when someone returns the stolen thing’ and ‘when my wife gives me manioc beer’; in general, when the breach that caused them to feel piinte is repaired. The causes of gane ponente as judged by the Waorani also fit fairly well with the core-relational theme that Lazarus (1991, p. 122) attributes to relief: “a distressing goal-incongruent condition that has changed for the better or gone away,” except that the “goal-incongruent condition that has changed” for the Waorani always seemed to have to do with the repair of a social breach.

Notice that many of the salient provocations of particular emotions for the Waorani have to do with cultural situations or items that are alien to members of most other societies. I imagine that I could ask 10,000 Americans what causes them to feel happy, and not one would offer I spear a wild pig with a lance. Nevertheless, Americans would be able to infer that it might constitute “making reasonable progress toward the realization of a goal” for the Waorani. Similarly, although the same 10,000 Americans might never offer I hear a jaguar howl in the forest as a reason to feel fright, they would be able to infer that it might constitute “facing an immediate, concrete, and overwhelming physical danger” for the Waorani.

To show this, I selected eight scenarios provoking each of five emotions that had been volunteered by my Waorani informants. The five emotions were tote ‘happy,’ ankai giñente ‘afraid,’ piinte ‘angry,’ gane ponente ‘calm/relief,’ and wæte ‘sad.’ The 40 scenarios provoking the five emotions are shown in Table 8. They were then presented to 202 American English native speakers (average age 26, 85 men and 117 women) with the following instructions:

The Waorani are slash-and-burn horticulturalists living in the tropical lowlands of Eastern Ecuador. We talked with a number of Waorani to find out what sorts of events cause them to feel various emotions.

22 For a more detailed discussion of Waorani ethnopsychology, especially as it relates to patterns of coalitional violence, see Boster et al. (2003).
Table 8  
Waorani emotion descriptors and the scenarios that were said to provoke them

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tote</strong> 'happy'</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>'I am talking with friends.'</td>
</tr>
<tr>
<td>8</td>
<td>'I spear a wild pig with a lance.'</td>
</tr>
<tr>
<td>12</td>
<td>'My father tells stories about the ancestors.'</td>
</tr>
<tr>
<td>15</td>
<td>'I am bathing in the river with my children.'</td>
</tr>
<tr>
<td>19</td>
<td>'I play soccer.'</td>
</tr>
<tr>
<td>28</td>
<td>'I am weaving hammocks or making baskets.'</td>
</tr>
<tr>
<td>32</td>
<td>'I do my work well.'</td>
</tr>
<tr>
<td>40</td>
<td>'I am with my family.'</td>
</tr>
<tr>
<td><strong>ankai giñente</strong> 'afraid'</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>'I see someone trying to kill me.'</td>
</tr>
<tr>
<td>4</td>
<td>'There is no food.'</td>
</tr>
<tr>
<td>9</td>
<td>'Someone in my family dies.'</td>
</tr>
<tr>
<td>13</td>
<td>'It is night and my children have not come home.'</td>
</tr>
<tr>
<td>16</td>
<td>'I see a boa constrictor on the path.'</td>
</tr>
<tr>
<td>18</td>
<td>'There is no water.'</td>
</tr>
<tr>
<td>20</td>
<td>'Other people are angry at me.'</td>
</tr>
<tr>
<td>21</td>
<td>'I hear a jaguar howl in the forest.'</td>
</tr>
<tr>
<td><strong>piinte</strong> 'angry'</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>'Somebody approaches my house without calling out.'</td>
</tr>
<tr>
<td>14</td>
<td>'My daughter elopes without my permission.'</td>
</tr>
<tr>
<td>24</td>
<td>'I come home from working and my wife doesn’t have the manioc beer ready.'</td>
</tr>
<tr>
<td>25</td>
<td>'I leave my manioc and plantains but they don’t bring the game meat they promised in exchange.'</td>
</tr>
<tr>
<td>31</td>
<td>'Someone doesn’t help on a community work project.'</td>
</tr>
<tr>
<td>33</td>
<td>'Others are spreading rumors about me.'</td>
</tr>
<tr>
<td>34</td>
<td>'Somebody borrows my canoe without permission.'</td>
</tr>
<tr>
<td>37</td>
<td>'My brother beats me up.'</td>
</tr>
<tr>
<td><strong>gane ponente</strong> 'calm/relief'</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>'The children are well behaved.'</td>
</tr>
<tr>
<td>11</td>
<td>'All my work is finished.'</td>
</tr>
<tr>
<td>22</td>
<td>'Someone takes something from me, but then returns it.'</td>
</tr>
<tr>
<td>23</td>
<td>'My child comes home after staying out late.'</td>
</tr>
<tr>
<td>26</td>
<td>'I am lying in my hammock at home.'</td>
</tr>
<tr>
<td>29</td>
<td>'I drink manioc beer.'</td>
</tr>
<tr>
<td>36</td>
<td>'I make up with someone I have fought with.'</td>
</tr>
<tr>
<td>39</td>
<td>'I am eating.'</td>
</tr>
<tr>
<td><strong>wæte</strong> 'sad'</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>'A tarantula bites me.'</td>
</tr>
<tr>
<td>3</td>
<td>'My mother died.'</td>
</tr>
<tr>
<td>6</td>
<td>'My house burns down.'</td>
</tr>
<tr>
<td>17</td>
<td>'My family doesn’t come.'</td>
</tr>
<tr>
<td>27</td>
<td>'My mother gets mad at me.'</td>
</tr>
<tr>
<td>30</td>
<td>'I don’t work.'</td>
</tr>
<tr>
<td>35</td>
<td>'My child dies.'</td>
</tr>
<tr>
<td>38</td>
<td>'I am whipped with stinging nettles.'</td>
</tr>
</tbody>
</table>
The following scenarios came out of those interviews – each of these events was judged by a Waorani to cause people to feel a certain emotion. We are interested in comparing their judgments with those of American English speakers. We want to know whether Americans will judge the emotional implications of the scenarios similarly to the way they are judged by the Waorani. Some of the scenarios involve things not encountered in US society (e.g., manioc beer and wild pigs), so you may have to use your imagination to figure out how some of these unfamiliar events would make you feel. It will take about 20 minutes to evaluate the 40 scenarios.

In any society, certain kinds of events are thought to provoke specific emotions. For each of the following scenarios, please write in the number or numbers of the emotion(s) from the following list that best describe how you would feel if the event had happened to you. (Often, the same event will cause you to feel a number of different ways and different events may cause you to feel the same way. Thus, you can enter more than one emotion number for each scenario and each emotion number could be applied to more than one scenario. However, you must enter at least one emotion number for each scenario.)

Each scenario was presented in the form “When [scenario], I feel _____,” as illustrated in Figure 18. The design of this task uses the mapping method, and is similar to that of the facial expression naming task, with the scenarios standing in for the photographs as the emotionally evocative stimuli and the responses limited to the 40 alternatives offered instead of being truly open-ended.

The results of the analysis of the data generated in this task are illustrated in Figure 19, Figure 19 is Plate 8.19 in the Separate Color Plate section. It shows a correspondence analysis of the mapping matrix of the 40 Waorani scenarios with the 40 English emotion terms. The numerical identifiers of the scenarios (shown in Table 8) are plotted along with the emotion terms the informants assigned to them. The figure captures at a glance some of the cultural similarities and differences between my Waorani and American informants.

On the one hand, it is clear that the Waorani emotional world is not “incommensurate” or impenetrable – American English native speakers are able to assign emotion terms consistently to the scenarios despite the fact that many are culturally alien. Most

![Table of Emotions](https://via.placeholder.com/150)

Enter the numbers of all the emotions that apply, separated by a comma.

<table>
<thead>
<tr>
<th>1 afraid</th>
<th>2 angry</th>
<th>3 annoyed</th>
<th>4 anxious</th>
<th>5 ashamed</th>
<th>6 astonished</th>
<th>7 bored</th>
<th>8 calm</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 confused</td>
<td>10 content</td>
<td>11 thoughtful</td>
<td>12 delighted</td>
<td>13 depressed</td>
<td>14 ecstatic</td>
<td>15 embarrassed</td>
<td>16 empathy</td>
</tr>
<tr>
<td>17 excited</td>
<td>18 frustrated</td>
<td>19 grief</td>
<td>20 happy</td>
<td>21 hate</td>
<td>22 hopeful</td>
<td>23 jealous</td>
<td>24 lonely</td>
</tr>
<tr>
<td>25 love</td>
<td>26 lust</td>
<td>27 nervous</td>
<td>28 proud</td>
<td>29 relaxed</td>
<td>30 sad</td>
<td>31 satisfied</td>
<td>32 secure</td>
</tr>
<tr>
<td>33 shocked</td>
<td>34 stressed</td>
<td>35 surprised</td>
<td>36 terrified</td>
<td>37 tired</td>
<td>38 upset</td>
<td>39 vengeful</td>
<td>40 secure</td>
</tr>
</tbody>
</table>

When I see someone trying to kill me, I feel

![Fig 18](https://via.placeholder.com/150)

Fig 18. Example of form used to elicit American English descriptors of scenarios derived from Waorani emotion research.
of the scenarios that cause the Waorani to feel tote ‘happy’ or gane ponente ‘calm/relief’ are judged by the Americans to provoke one to feel excited, ecstatic, hopeful, happy, delighted, content, calm, secure, proud, relaxed, love, and satisfied. Most of those that cause the Waorani to feel ankai giñente ‘afraid’ are judged by the Americans to provoke one to feel terrified, afraid, nervous, worried, anxious, surprised, stressed, astonished, or shocked; and so on. On the other hand, the American informants do not always agree with the Waorani in their appraisals of the scenarios. Being bitten by a tarantula (scenario 2) was volunteered by my Waorani informants as something that would make one feel wæte ‘sad,’ but was usually judged by the Americans to make one afraid. Someone approaching one’s house without calling out (scenario 7) was volunteered by the Waorani as something that would make one feel piinte ‘angry’ but was judged by the American to make one anxious, and so on.

Part of the apparent difference here is probably due to the fact that human reactions to events are almost certainly universally complex. Renato Rosaldo’s Ilongot informants are not alone in feeling a mixture of grief and rage at the death of kin [Rosaldo (1989)] – the

![Fig. 19. Correspondence analysis of American English emotion descriptors and Waorani emotion scenarios.](image-url)
Waorani feel that and fear as well. Thus scenario 9 (‘someone in my family dies’), which was volunteered as something that would make Waorani feel ankai giñente ‘afraid,’ is very similar to scenarios 3 (‘my mother died’) and 35 (‘my child dies’), which were volunteered as things that would make Waorani feel wæte and all three are judged by the Americans to provoke one to feel sad, lonely, depressed, and grief. My sense is that Americans also often feel rage and fear in addition to grief at the loss of a loved one, but may be less likely to acknowledge it than the Waorani or the Ilongot. However, many of the differences are probably due to genuine differences in how the various scenarios are appraised by Waorani and American informants and by differences in the meanings of the terms in the emotion lexicons that the two groups have at their disposal.

This experiment leaves considerable room for improvement; a better research design would have presented the same scenarios in the same format to be judged by both Waorani and Americans. Unfortunately, it is virtually impossible to administer the paper-and-pencil tasks so familiar to American undergraduates among the Waorani, especially to older non-literate monolinguals. Nevertheless, as expected, the experiment shows that even scenarios that are culturally unfamiliar to American informants (e.g., ‘I spear a wild pig with a lance,’ ‘I hear a jaguar howl in the forest’) are appraised in terms that would sound reasonable to the Waorani (i.e., delighted and terrified, respectively). This means that, with care to choose scenarios for comparison that represent more universal human possibilities (e.g., ‘my child comes home after staying out late,’ ‘my mother died,’ ‘others are spreading rumors about me’), one could design an instrument expressed in something close to Wierzbicka’s semantic metalanguage that could be appraised by speakers of a number of different languages.

5. Conclusion

To conclude, I would like to outline a research design for carrying out the next step in exploring the pattern of similarity and difference among cultural emotion systems.

First, the design should use the method of mapping, presenting a collection of emotionally evocative objects or propositions to informants to be evaluated. Representations of facial expressions can be included among the stimuli presented, but the investigation should not be limited to decontextualized, disembodied faces. Whatever collection of stimuli is chosen, it should have, as far as possible, a universal emotional import. For example, if exploring causes of emotional states, it is best to choose scenarios that can happen anywhere (e.g., ‘my child died,’ ‘my lover left me for another’) over scenarios that are only familiar to a minority of societies (e.g., ‘I got an A on my organic chemistry final,’ ‘I spear a wild pig with a lance’). The stimuli should be chosen with an eye to capturing the sorts of distinctions between emotion terms made in the ethnographies of emotion we already have in hand [e.g., Briggs (1970), Levy (1973), Rosaldo (1980), Lutz (1988), Heider (1991)]. In other words, the sort of situations that are said to provoke lek, schadenfreude, fago, amae, popokl, and liget should be included in any set of propositional stimuli. Any propositions should be expressed in something approaching...
Wierzbicka’s (1992) semantic metalanguage, employing simple, clear terms with universal (or near universal) significance.

The design should explore the connection of emotional responses to the entire model of the mind [D’Andrade (1987)]. In addition to discovering what scenarios are most likely to provoke various emotional responses, we would also want to find out (at a minimum): (1) what external signs or manifestations show that someone is feeling each emotion; (2) what internal states and sensations are associated with each emotion; (3) what wishes, desires, and goals are implicated in the emotional response; (4) what thoughts and perceptions are linked with each emotion; (5) what actions individuals are likely to take if they experience each emotion; (6) what responses others are likely to take on seeing the emotional display; (7) what types of personalities are most likely to experience each emotion; and (8) what is the moral and/or religious interpretation of each emotion.

By “each emotion,” I would advocate as generous a span of emotional responses or states as the language allows. Some investigators of emotion have attempted to delimit their attention to emotions proper, rather than have it spread over a host of related phenomena. Thus, true emotions are distinguished from reflexes by the amount of cognitive appraisal involved (e.g., surprise vs. startle); are distinguished from moods and personality traits by their short duration and the specificity of their causation (e.g., irritated vs. grumpy, paranoid); and are distinguished from sensations of bodily states by their lesser degree of corporeality (e.g., anger, sadness, happiness, fear vs. hunger, sleepiness, pain, thirst). Furthermore, within any family of related emotions, one can distinguish grades of intensity (e.g., annoyance, irritation, anger, rage, fury); “simple” emotions from combinations (e.g., sadness or anger vs. nostalgia, “love-hate,” anxiety, etc.); and the number of obligate actors or objects involved in the emotion (e.g., ‘I am sad,’ ‘I love you,’ ‘I am jealous of your flirting with him’). Unfortunately, it is hard enough to keep these distinctions straight in English, let alone to expect them to be universally relevant. The better alternative is to explore as wide a variety of terms as possible; in English, this amounts to nearly everything that fits the frame I feel _____.

Because this program involves showing a very large array of emotionally evocative stimuli and allowing informants to respond with a very wide array of emotional descriptors, a balanced block design should be employed to cut down on the amount of the whole instrument any one informant would have to respond to.

Finally, this systematic elicitation should be complemented with a deep ethnography of the ethnopsychology of the group, so as to ensure that themes or concerns of local importance are captured in the systematic instrument.

References

Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley, CA).


Ch. 8: Emotion Categories Across Languages


Chapter 9

THE WORLD COLOR SURVEY DATABASE

RICHARD S. COOK
University of California, Berkeley

PAUL KAY
International Computer Science Institute, University of California, Berkeley, California

TERRY REGIER
University of Chicago

Contents

Abstract 224
1. Introduction 225
2. The WCS: History and methodology 225
3. Data processing and analysis 228
4. Cleaning the data 232
5. Original format of the data and creation of the WCS Online Data Archive 232
6. Uses of the WCS archive 234
   6.1. Universals of color naming 234
   6.2. Variation in color naming 237
7. Conclusion 240
References 240

1 Corresponding author: kay@cogsci.berkeley.edu. International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA.
Abstract

The World Color Survey (WCS) is a research project that was undertaken to validate, invalidate, or – most likely – modify the main findings of Berlin and Kay (1969) (B&K): (1) that there exist universal crosslinguistic constraints on color naming, and (2) that basic color terminology systems tend to develop in a partially fixed order. To this end, the WCS collected color naming data from speakers of 110 unwritten languages. The data have recently been compiled into a unified data archive, available online at http://www.icsi.berkeley.edu/wcs/data.html. In this chapter, we review the history of the WCS, including the creation of the online data archive, and describe our recent use of the archive to test the universality of color naming across languages.
1. Introduction

The World Color Survey (WCS) is a research project that was undertaken to validate, invalidate, or – most likely – modify the main findings of Berlin and Kay (1969) (B&K): (1) that there exist universal crosslinguistic constraints on color naming, and (2) that basic color terminology systems tend to develop in a partially fixed order. To this end, the WCS collected color naming data from speakers of 110 unwritten languages. The data have recently been compiled into a unified data archive, available online at http://www.icsi.berkeley.edu/wcs/data.html. In this chapter, we review the history of the WCS, including the creation of the online data archive, and describe our recent use of the archive to test the universality of color naming across languages. Section 2 recounts the general history and methodology of the WCS, Section 3 deals with data processing and analysis. Section 4 covers data checking and quality control procedures. Section 5 describes the original format of the data and steps taken to create the online database. In Section 6, we discuss the use of the online database to test two hypotheses: one involving universal tendencies in cross-language color naming and the other dealing with explanation of a particularly salient instance of variation in cross-language color naming. Section 7 constitutes a brief conclusion.

2. The WCS: History and methodology

The WCS was begun in 1976 to check and expand B&K’s findings in a full-scale field study. B&K had investigated the color terminology systems of 20 languages in the following way. The stimulus array used by Lenneberg and Roberts (1956), consisting of 320 Munsell chips of 40 equally spaced hues and eight levels of lightness (Value) at maximum saturation (Chroma) for each (Hue, Value) pair, was supplemented by nine Munsell achromatic chips (black through gray to white) – the resulting stimulus array is shown in Figure 1a. First, without the stimulus array present, the major color terms of the collaborator’s native language were elicited by questioning that was designed to find the smallest number of simple words with which the speaker could name any color (basic color terms). Once this set of basic color terms was established, the collaborator was asked to perform two tasks. In the naming task, the stimulus array was placed before the speaker and for each color term \( t \), a piece of clear acetate was placed over the stimulus board and the collaborator was asked to indicate, with a grease pencil on the acetate sheet, all the chips that he or she could call \( t \). In the focus task, the stimulus array was shown as before and the collaborator was asked to indicate the best example(s) of \( t \) for each basic color term \( t \). The boundaries of categories showed great variability, perhaps because of the

---

2 Actually, Figure 1a shows the very slightly modified stimulus palette used in the WCS. The B&K stimulus array lacked achromatic chip A0.

3 For a review of the B&K basicness criteria, as well as of the notion of basic color terms in B&K and other literature, see Maffi (1990).
Fig. 1  (a) The WCS stimulus array.  (b) Munsell and WCS coordinates for stimulus palette.  The leftmost column and the top row give the WCS coordinates for lightness and hue, respectively.  The rightmost column and the bottom two rows give the Munsell coordinates for Value and Hue, respectively.  Entries in the body of the table show the corresponding Munsell Chroma numbers.  [With regard to the A and J rows, there are no Munsell hues at the extremes of Value (lightness): 9.5 (white) and 1.5 (black).]
vagueness of the instruction for the naming task: probably some subjects took the instruction to call for all the chips that were more than anything else, while others appear to have taken it to call for all chips in which any trace of was visible. The focal choices of the B&K subjects were much more clustered and led to the conclusion that

... [1] the referents for the basic color terms of all languages appear to be drawn from a set of eleven universal perceptual categories, and [2] these categories become encoded in the history of a given language in a partially fixed order. Berlin and Kay (1969, pp. 4–5).

In retrospect, the B&K study – only 20 languages directly assessed with calibrated color stimuli and all the work done in the San Francisco Bay Area – can be viewed as a pilot project for the WCS. The B&K results were immediately challenged, mainly by anthropologists, on the grounds that the sample of experimental languages was too small, too few collaborators per language were questioned (sometimes only one), all native collaborators also spoke English, the data were collected in the San Francisco Bay area rather than in the homelands of the target languages, certain regions of the world and language families were underrepresented or overrepresented in the sample of 20, and the sample of 20 had too few unwritten languages of low-technology cultures [Hickerson (1971), Durbin (1972), Collier (1973), Conklin (1973)]. The results were nevertheless supported by various ethnographic and experimental studies conducted after 19697 and were largely accepted by psychologists and vision researchers [e.g., Brown (1976), Miller and Johnson-Laird (1976), Ratliff (1976). See also Boynton (1997, p. 133 ff), Kaiser and Boynton (1996, p. 498 ff).

In the late 1970s, through the cooperation of SIL International (then the Summer Institute of Linguistics), which maintains a network of linguist-missionaries around the world, data on the basic color term systems of speakers of 110 unwritten languages representing 45 different families and several major linguistic stocks were gathered in situ. Fieldworkers were provided with a kit containing the stimulus materials (330 individual chips in glass 35-mm slide sleeves for the naming task and the full stimulus board for the focus task) as well as coding sheets on which to record collaborators’ responses. The

---

4 MacLaury later demonstrated that speakers can often be induced to increase the number of chips they will indicate as belonging to a given term simply by asking them if there are “any more”; speakers frequently increase the size of a named category several times in response to this “mapping” task [MacLaury (1997, pp. 77–84 et passim)].

5 B&K extended their findings on the 20 languages assessed experimentally to another 78 reports of color terminology systems they found in the literature.

6 Initial support for the WCS was in the form of NSF grant BNS 76-14153. Subsequent NSF support was furnished by grants BNS 78-18303, BNS 80-06808, SBR 94-19702, BCS 01-30420, and BCS 04-18283. NSF support of the WCS project is gratefully acknowledged, as is additional support by the University of California, Berkeley, the Summer Institute of Linguistics (now SIL International), and the International Computer Science Institute. We would also like to express our most sincere gratitude to the many field linguists of the SIL who unselfishly devoted long hours to what for many must often have been an unwelcome task.

instructions requested that fieldworkers collect data from at least 25 speakers, both males and females, and urged them to seek out monolingual speakers insofar as possible. The modal number of speakers actually assessed per language was 25 and the mean number was 24. (A facsimile of the WCS instructions to fieldworkers and of the original coding sheets is available at http://www.icsi.berkeley.edu/wcs/images/WCS_instructions-20041018/jpg/border/index.html.) The aim was to obtain names, category extent and best examples of basic color terms in each language – basic color terms being described in the instructions as “the smallest set of simple words with which the speaker can name any color.”

The WCS methodology coincided with that of the B&K study in the use of a standardized set of Munsell color chips, consisting of 320 chromatic chips representing 40 equally spaced hues at eight levels of lightness (Munsell Value), each at maximum available saturation (Munsell Chroma). One white chip was added in the WCS study that was whiter than any chip available at the time of the B&K study, making for a total of 10 achromatic chips and an overall total of 330 chips, as shown in Figure 1a. The Munsell notations of the chips employed and the simplified notation used for precisely this palette by the WCS project are shown in Figure 1b.

The WCS differed from B&K in the technique for eliciting naming responses. In the WCS procedure, no preliminary interview was administered to establish a set of basic color terms, and in the naming task the 330 individual color stimuli were shown to each cooperating speaker, one by one, according to a fixed random order, and a name elicited for each (in contrast with the B&K procedure of presenting the entire stimulus array at once to elicit naming responses). Fieldworkers were instructed to urge observers to respond with short names (although, depending on the morphology of the language, particular field circumstances, and local culture, there was considerable variation in the degree to which the field investigators were able to satisfy these desiderata). Identification of basic color terms, therefore, was done by the fieldworker as a result of the naming task itself, rather than through prior elicitation. The best example (focus) responses were elicited in the same way in both studies: once a set of basic color terms was isolated, the native observer was presented with the full palette (in WCS, a physically improved version of the original Munsell chip board, devised by Collier et al. 1976) and asked to indicate the chip (or chips) that represented the best example of each term, one by one.

3. Data processing and analysis

Once data gathering was completed (Ca. 1980), data processing, quality control, and analysis were undertaken at University of California, Berkeley, and at SIL in Dallas.

Computer programs were developed for both data entry and data analysis. The original processing yielded, for each language, a preliminary data summary that included the following information:

* Language name and location.
* Name, age, sex, and other vital statistics of each speaker interviewed.
• List of terms used, each with a tentative gloss and a typographical symbol representing it in the naming and focus arrays. (See Figure 2a for an example. All the examples in Figure 2 are for the Niger-Congo language Wobé of Côte d’Ivoire. The information shown in Figure 2 is not that of the initial data entry and preliminary processing but of the final results of checking, following corrections to the original data entry and preliminary analyses, as described below.)

• Individual naming arrays, structured by the form of the full stimulus array shown in Figure 1 and presenting, for each speaker, the full picture of his or her use of color terms from the naming task (see Figure 2b for examples).

• Individual focus arrays, presenting, for each speaker, the full picture of his/her focal (best example) choices from the focus task. (see Figure 2c for examples).

• Aggregate naming arrays, also in the form of the stimulus array, presenting the aggregated results of the naming task across all speakers, at various levels of interspeaker agreement. For example, for a language with 25 native observers, the 40% naming aggregate shows for each stimulus chip c (1) the symbol for the most popular name given to c, if at least 10 speakers gave c that name or (2) a blank if no single name was given to c by 10 or more speakers. (see Figure 2d for examples).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Term</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>kpe’</td>
<td>black/green/blue</td>
</tr>
<tr>
<td>+</td>
<td>“plua-”</td>
<td>white</td>
</tr>
<tr>
<td>#</td>
<td>-sain’</td>
<td>red/yellow</td>
</tr>
</tbody>
</table>

(Diacritics stand for tones: ‘ = very high; ‘ = high; - = low.)

---

(a) (b)

Fig. 2 (a) Wobé basic color terms, with glosses and key symbols. (b) Naming responses for four Wobé speakers (see previous figure for terms denoted by typographical symbols).
Subsequently, an additional kind of array was produced, called a term map. A term map for a given term furnishes a visual picture of the relative frequency of that term’s usage over the stimulus space in the form of a kind of contour map. A term map is thus a display of the denotation of a color term. Conceptually, one can imagine a 3D histogram for a term \( t \) in which the stimulus surface constitutes the floor plane and the height of the column over each stimulus chip \( c \) represents the proportion of speakers using \( t \) to name \( c \). We represent two dimensionally a contour map of such a 3D histogram, viewed from above: this is what we call a term map.

Specifically, a term map for a term \( t \) is a display of the stimulus surface where the symbol appearing on chip \( c \) is

- #, if 81% or more speakers who used \( t \) named \( c \) with \( t \),
- +, if 61 – 80% of the speakers who used \( t \) named \( c \) with \( t \),

Fig. 2 (c) Focus (best example) responses for four Wobé speakers. (d) Aggregate naming arrays for 25 Wobé speakers. (Note that at the 40% level of agreement all 330 chips were named. That is, at least 10 speakers gave the modal response for each of the 330 chips. Wobé was a high-consensus language.)
• –, if 41 – 60% of the speakers who used t named c with t,
• -, if 21 – 40% of the speakers who used t named c with t,
• nothing (blank), if 20% or fewer of the speakers who used t named c with t,
• @, if the percentage of t-users who used t to name c equaled or exceeded the percentage of t-use by t-users for any other chip. The numerical value of @, the consensus level for t, is given at the bottom of each term map, as well as the number of collaborating speakers of the language and the number using the term mapped.

The density of the symbols as visual objects increases as the proportion of respondents they represent increases; thus, a term map gives a somewhat iconic representation of a term as a gradient category, where proportion of speakers using t to name c is taken as a proxy for the degree of membership of color c in the gradient category named by term t.\(^8\)

Figure 3 shows the term maps for Wobé.

---

8 The conventions for representing the various displays in Figures 2 and 3 were developed in an age of typewriter technology.
4. Cleaning the data

Before the initial data entry had been checked, a set of microfiche summaries containing preliminary versions of the data arrays described above was accidentally made available to the public in 1991. This release of unchecked data was unfortunate because we subsequently discovered that these summaries contained many errors, especially in the assigning of similar spellings to same or different terms. Coders had frequently made snap judgments about probable morphological variations in unfamiliar languages. Such errors of primary interpretation, along with simple input errors – using the same abbreviation for two different terms, using two different abbreviations for a single term, and the like, introduced significant inaccuracy into the data in the 1991 microfiche summaries. The current archived data are based on a number of data-cleaning procedures subsequently adopted: (1) checking the electronic record against the original paper coding booklets, (2) carefully surveying fieldworkers’ notes regarding spelling variations and morphological structure, (3) rerunning various summary programs to see if, for example, two distinct “terms” with similar spellings had virtually identical term maps (indicating mere spelling variation), (5) corresponding with the original investigators in some cases and with other specialists in the same or related languages in other cases regarding morphological analysis of the recorded color terms, and (6) in general checking all possible information from whatever source to make as certain as possible that what our data listed as the roster of color terms of a language was an accurate rendition of the color terms of that language as recorded by the field investigators.

5. Original format of the data and creation of the WCS Online Data Archive

The early work of converting the handwritten coding sheets (prepared in the field) to electronic format was split between a team of researchers at SIL in Dallas, Texas, and a team at UC Berkeley. The two halves were eventually joined together when the SIL data came to Berkeley in the mid-1980s. At that point, the WCS data were not stored in a single database but rather in 110 separate directories, one for each language. Within a language directory, the data were stored in four main files (with some subsidiary files): (1) an Informants file, containing the name and vital statistics of each native collaborator for that language, along with an identification number and some ancillary information regarding other languages spoken, etc.; (2) a Dictionary file, containing the color terms of the language and one or more abbreviations with which they could be referred to in other files; (3) a Naming data file, containing the naming data for each collaborating speaker; and (4) a Focus file, containing the best example(s) of data for each speaker. The WCS data remained in these 110 separate and unlinked directories until December 2000.

The primary copies of the World Color Survey electronic computer files, described above, were housed at UC Berkeley on an aging hard disk connected to an old computer, with accompanying fragile removable media backup disks. In December 2000,
those files were all compressed into a single archive, copied to a campus server, and then burned to CD-ROM. In early 2002, work began to extract the data from that archive, to create an operating system-independent, public, online archive. We describe the archive creation process here, so that users of the archive may have a clear picture of the nature of the data and their origin.

Initial study of the electronic source files revealed that a number of components in different formats would have to be processed separately for integration into a coherent whole. A major reason for this was that the SIL and Berkeley teams had applied different conventions to the digitization of their respective data.

The results of the naming task (presented in the current online archive in the file “term.txt”) constituted by far the bulk of the data, totaling nearly 1 million lines of text. These naming data, in 110 separate files, had received the greatest attention of all the existing files, and by the year 2000 had attained a fairly stable state. Other components of the data included focus data (from the best-example task), speaker data (personal information on the native observers), language data (geographical information on the languages), dictionary data, and analyses of the color terms present for each language. Compared to the naming data, each of these latter components had received considerably less attention over the years. Unlike the naming data, the separate Berkeley and Dallas focus data had never been consolidated into a coherent whole. Instead, portions of each had been partially processed, and in 2002 we began to reassemble the pieces and complete the processing.

The source files fell into two major groups, labeled “new” and “old.” The “new” files represented the effort to combine the Berkeley and Dallas data. The old files themselves were still split into “Berkeley” and “Dallas” groups. There were omissions in the “new” files of focus data only available in the “old” files, and there was a certain amount of overlap and variability in formatting among all three. The task then became one of identifying omissions in the “new” focus data, filling in these gaps, and verifying the new focus data where they were present. Gaps in the new data were filled in from the primary electronic source of the missing data (Berkeley or Dallas). In the event that no electronic data were available, it was necessary to revert the coding sheets and input the data afresh, interpreting the fieldworkers’ conventions as best these could be determined. Fortunately, recourse to the original coding sheets was necessary only for a relatively small number of languages.

The data from the coding sheets had originally been entered into the computer systems via a process that involved creating an electronic “dictionary,” that is, an inventory of the terms attested on all coding sheets for all speakers of the language. A unique abbreviation (WCS code) was then assigned to each of the terms, often bearing an obvious relation to those used by the original fieldworker. Different fieldworkers had used different transcriptional and notational conventions. Thus, the electronic Dictionary data contained both an ASCII (or other encoded) interpretation of the fieldworker’s transcription of the native color term, and a one- or two-letter abbreviation to be used in the input of both naming and focus data. These three repositories of term abbreviations had to ultimately agree among themselves, and also to be transparently related to
the rendering of the terms on the coding sheets. Where there was disagreement, a correction had to be applied, based on assessment of all of the available documentation.

The original fieldworkers themselves had employed various collection and organizational techniques, all of which contributed another level of variability to the data. In examination of the coding sheets, it became apparent that some fieldworkers were fastidious workers who had clearly put great effort into selecting their native collaborators, collecting the responses, and copying out the final coding sheets in pen or with a typewriter, using IPA for their transcriptions. At the other extreme, data on the coding sheets were sometimes barely legible and sometimes internally inconsistent, as when the same abbreviation was used for evidently different terms or two different abbreviations were used for the same term. Once the relations among the various data formats became clear, we compiled the data together into a single data archive composed of four files in tab-delimited plain text format: one containing the naming data (“term.txt”), one containing best-example data (“foci.txt”), one describing the languages (“lang.txt”), and one describing the individual speakers of the languages (“spkr.txt”). The various “dictionary” files have also been combined into a single file compatible with the other four, and integrated into the relational database system. This dictionary data appears in the online archives in UTF-8 format under the name “dict.txt.” Online documentation concerning the formats of these files is included with the archive. The work to prepare the data consumed all of 2002, and it was not until January 2003 that the first portion of the online data was released to the public.

6. Uses of the WCS archive

The WCS data archive has been used in investigating two broad questions, one concerning universals and the other concerning variation in color naming.

6.1. Universals of color naming

Since B&K found evidence for universals in color naming across languages, the existence of such constraints has generally been accepted in the scientific community. However, there have always been dissenters from this consensus [e.g., Durbin (1972), Hickerson (1971)], and this dissenting view has recently gained some prominence [e.g., Davidoff, Davies, and Roberson (1999) Lucy (1992, 1996, 1997), Roberson, Davies, and Davidoff (2000), Saunders and van Brakel (1997)]. Criticisms of the universalist position have come in two major varieties. The first points out that B&K’s findings were never objectively tested, as they relied on visual inspection of color naming data. Lucy (1997) challenges such a methodology as hopelessly subjective:

[Work in the B&K tradition] not only seeks universals, but sets up a procedure which guarantees both their discovery and their form. … when a category is identified ... it is really the investigator who decides which “color” it will count as … What appears to be objective – in this case, a statement of statistical odds – is [not]. (p. 334)
In this view, B&K’s subjective methodology allowed them to impose their own universalist assumptions on their data – so the universals are actually in the minds of the investigators, not in the languages of the world. The second strand of criticism points out that B&K’s data were drawn primarily from written languages, and thus may not be representative. This point is coupled with analyses of particular unwritten languages, which are claimed to counterexemplify universal constraints [e.g., Berinmo: Davidoff et al. (1999), Roberson et al. (2000), Hanunóo and Zuni: Lucy (1997)]. Kay (1999) has responded to this with counter-analyses of these languages, arguing that each fits neatly into the universal pattern. Disputes of this sort over conflicting interpretations of individual color naming systems could continue indefinitely without resolving the main issue of whether universal, cross-language constraints on color naming systems actually exist. We wished to resolve this issue in a manner that would respond to both varieties of criticism.

To that end, Kay and Regier (2003) used the WCS database and the B&K data to objectively test the hypothesis that color terms across languages cluster together more tightly in color space than would be expected by chance. This was done as follows. First, for each term in each WCS language, the centroid (i.e., center of mass) of each speaker’s naming distribution was calculated, after translation of Munsell coordinates into CIE L*a*b* coordinates [Wyszecki and Stiles (1967)]9. We refer to the resulting point as the “speaker centroid” for that speaker and that term. For each term, we then calculated the “term centroid”: the centroid of the speaker centroids for that term. This produced a point representation of the term in CIE L*a*b* space. Each centroid was coerced to the nearest point in the stimulus array, so that our point representation of the term resided within the set of points out of which it was constructed. The speaker centroids were plotted over the stimulus space, yielding the picture shown in Figure 4a, Figure 4a is Plate 9.4a in the Separate Color Plate section. Intuitively, the speaker centroids are not distributed randomly or evenly over the stimulus space. Figure 4a shows sharp peaks and broad valleys in the distribution of speaker centroids. We showed this clustering of speaker centroids to be statistically significant by the results of a Monte Carlo simulation on the term centroids, depicted in Figure 4b – this time demonstrating universality across color terms without regard to their frequency of use, since each term is now represented by one centroid, whatever its frequency. In that simulation, a measure of dispersion (the opposite of clustering) of term centroids was defined as the sum, across languages, of the distances between each term (centroid) in that language and the closest term (centroid) in another language. This measure was calculated in the WCS data and in 1000 hypothetical randomized datasets, each created by rotating the actual WCS naming centroid distribution a random degree of hue angle in CIE L*a*b* space (to maintain the shape of the distribution while randomizing its location in perceptual space). In Figure 4b, Figure 4b is Plate 9.4b in the Separate Color Plate section it can be seen that the dispersion measure for the actual WCS dataset falls well below the lower bound of the distribution of 1,000 hypothetical WCS datasets.

9 CIE L*a*b* is a three-dimensional color space; its creators made a systematic effort to assure that local Euclidean distance corresponds to perceptual dissimilarity.
datasets, indicating a probability less than 0.001 that the degree of clustering in the real WCS dataset is the result of chance. A similar Monte Carlo test also revealed that color terms in the unwritten languages of the WCS tend to cluster near the color terms of written languages (Berlin and Kay 1969). (For further explanation, see Kay and Regier 2003.)

Fig. 4 (a) Contour plot of WCS speakers’ naming centroids, compared with English naming centroids (black dots). (Source for English naming centroids: Sturges and Whitfield 1995.) The outermost contour represents a height of 100 centroids, and each subsequent contour represents an increment in height of 100 centroids. Source: Kay and Regier (2003). (b) Monte Carlo test for clustering within the WCS data. The distribution of dispersion values shown in blue was obtained from 1000 randomized datasets. The red arrow indicates the dispersion value obtained from the WCS data. [Source: Kay and Regier (2003)].
The above results concern the naming data from the WCS. Statistical tests of the degree of clustering of WCS best-example (focus) choices remain to be performed, but a preliminary plot of the focus data suggests strongly that such tests will turn out to be statistically significant. (Black dots represent the naming centroids for the English terms indicated.)

In the WCS focus distribution, the chips receiving the highest numbers of focus choices were J0 (black) and A0 (white), not shown in Figure 5, Figure 5 is Plate 9.5 in the Separate Color Plate section. In Figure 5, restricted to the WCS chromatic chips, two of the four major focus peaks fall, one each, on the English yellow and English green naming centroids. A third WCS focus peak falls on a chip adjacent to the English red naming centroid and the fourth major peak in the WCS best-example distribution falls two chips away from the English blue naming centroid. These observations suggest strongly that objective tests will show a non-chance association between the highest peaks of the WCS focus distribution and points in color space favored by English color naming.\footnote{Note added in proof: Recently, such objective tests have yielded confirmatory results (Regier, Kay and Cook 2005).}

6.2. Variation in color naming

The above results demonstrate universal constraints in color naming. Yet there is also considerable cross-language variation, and it is still an open question why languages vary as they do in the naming of colors.

Fig. 5. Contour plot of WCS chromatic focus peaks compared with English naming centroids. [Source for English naming centroids: Sturges and Whitfield (1995)].
Lindsey and Brown (2002) provided a provocative answer for one aspect of this question, the investigation of which has employed the WCS online data archive. Some languages have separate terms for blue and green, while others have compound green-or-blue (“grue”) terms; Lindsey and Brown asked why this should be. They suggested that grue terms may derive from a sunlight-induced yellowing of the ocular lens: with a yellowed lens, short wavelengths are disproportionately filtered and blue stimuli would appear green, and would be named by the word for green. In other words, grue terms in this view are really words for green, and they extend to what normal eyes see as blue only because the yellowed lens distorts the perception of color. In support of this hypothesis, Lindsey and Brown noted that the proportion of languages with grue terms (rather than separate green and blue terms) is well predicted by the amount of UVB radiation from sunlight that strikes the earth’s surface where those languages are spoken – as would be predicted if grue is ultimately traceable to sunlight-induced lens yellowing. They also showed that speakers of English who were shown stimuli that artificially simulated sunlight-induced yellowing of the lens extended the English word green to include stimuli presenting a spectral distribution comparable to that of a blue stimulus viewed through a yellow filter.

Intriguingly, Lindsey and Brown note that their hypothesis has the potential to “explain away” some recent findings suggesting a Whorfian influence of language on color cognition. Davidoff et al. (1999) examined color naming and memory in speakers of Berinmo, a language that has an enlarged yellow term, extending into the region that would be named “green” in English; this enlarged yellow category shares a border with a grue category. Davidoff et al. examined how well Berinmo speakers remembered colors straddling the boundary between these two Berinmo categories, and found that their performance was better for these colors than it was for colors straddling the boundary between English yellow and green. English speakers showed the opposite pattern. These findings suggest that a language’s color terms may influence color cognition for speakers of that language [see also Kay and Kempton (1984)]. However, Lindsey and Brown suggested a different interpretation of these data. They argued that since Berinmo has a grue term, its speakers may have yellowed lenses. This would explain why Berinmo’s yellow term expands into green: because yellowish greens are seen as more yellow through this lens. And it would also explain why color memory covaries with color naming across English and Berinmo: because both memory and naming are shaped by color perception, and that perception may be distorted by a yellowed lens in Berinmo speakers, relative to English speakers.

Regier and Kay (2004) tested this lens-yellowing hypothesis further, by probing a prediction it makes concerning the best examples of grue categories. If grue is really a green category that extends into blue because of a distorted perceptual color space, then there should be a single peak in the best example choices for grue (since there is a single peak for green), and it should fall somewhere between focal green and focal blue. However, if grue is instead a genuine abstraction over green and blue in an undistorted perceptual color space, the best examples for grue should peak either at green, or at blue, or at both. This prediction was easily tested using the focus data from the WCS
data archive. We found that best examples for grue terms peak at English green and very near English blue [Figure 6, Figure 6 is Plate 9.6 in the Separate Color Plate section; see also MacLaury (1997, pp. 234–235), compare also with Figure 5, which shows focal choices from all WCS languages combined, including those that distinguish between green and blue]. This suggests that the lens-yellowing hypothesis is incorrect. In doing so, it also indirectly supports the Whorfian hypothesis, by removing a competing explanation for the findings from Berinmo.

In response, Lindsey and Brown (2004) have presented further analyses of WCS focus data. Their findings confirm that grue best-example choices tend to peak at focal green and focal blue – which argues against their hypothesis. But at the same time, their findings also leave open the possibility that there may be a subset of speakers whose best-example choices fall between universal focal green and blue, and thus are consistent with their hypothesis. Further analysis will be needed to establish whether the observation of a small number focal choices between green and blue provides support for the Lindsey and Brown UVB hypothesis. In any case, the hypothesis seems imperiled on other grounds as well. A recent finding has shown that color naming in English by individuals with naturally yellowed ocular lenses does not differ from that by individuals with nonyellowed lenses [Hardy, Frederick, Kay, and Werner (2005)]. The probable reason for the difference observed between Lindsey and Brown’s simulation of yellowed optical media and naturally yellowed optical media is that in the latter case long-term processes of adaptation have time to operate [Delahunt et al. (in press), Neitz et al. (2002), Schefrin and Werner (1990, 1993)], in order to perceptually compensate for the increased yellowing of the lens. This makes it quite unlikely that lens yellowing could account for grue terms in the world’s languages.

Fig. 6. Contour plot showing the distribution, over chromatic stimuli, of best examples of grue terms in the WCS. Outermost contour represents a height of 10 hits; each subsequent inner contour represents a height increment of 10 hits. [Source: Regier and Kay (2004)].
7. Conclusion

The WCS data archives are a publicly accessible resource, available to all who wish to pursue questions related to color categorization across languages. We have provided this background to orient potential users of the archive – to give them a sense for where the data came from, how the data were compiled into an archive, and what sorts of questions the data can be used to investigate. We hope the archive proves to be a useful and flexible research tool for the scientific community as a whole.

References

Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley and Los Angeles)
Abstract

This chapter uses the concept of an [ATOM] to discuss ways in which categorization depends on conceptual change. Features of such theoretical concepts are theory-generated rather than observed. They fit much better with the “knowledge approach” to the nature of concepts than to classical, exemplar, or prototype theories. The concept of atom has undergone numerous changes in the history of chemistry, most notably the realization that atoms are divisible and have internal structure.
1. Introduction

Conceptual change is an important topic in many areas of cognitive science, including philosophy of science, developmental psychology, and science education. This chapter will discuss the relevance of cognitive processes of conceptual change to questions about categorization, particularly concerning the role of theoretical concepts such as atom, molecule, and element.

First some terminological clarification is in order. Philosophers such as Aristotle and Kant have used the term “category” to refer to the most fundamental concepts such as substance, quantity, and quality. We shall, however, follow the standard usage in psychology, in which a category is a class of things in the world, for example the class of dogs. Concepts are mental representations that usually correspond to particular words and refer to classes of things in the world. For example, the concept [DOG] corresponds to the word dog and refers to the class of dogs in the world. Categorization is the process of dividing the world into categories, and usually involves constructing concepts that provide mental representations of those categories.

Most psychological research on categorization involves classes of observable objects such as animals and furniture. In contrast, most philosophical discussion of conceptual change has concerned classes of non-observable objects such as atoms, black holes, and genes. Such objects are called theoretical entities, and concepts that refer to them are theoretical concepts. A crucial part of scientific inquiry involves the formation and improvement of such concepts, which categorize non-observable objects. Categorization is therefore more than dividing up the world based on the observed features of things; in fact, it requires one to create concepts that enable deep explanations of how the world appears by hypothesizing properties of non-observable entities. For example, the ancient Greeks formulated the concept of fundamental particles called atoms in order to explain many facts about the natural world.

The development of the concept [ATOM], with the attendant development of concepts such as [MOLECULE] and [ELEMENT], carries numerous lessons for cognitive theories of categorization and conceptual change. We will argue for the following conclusions:
1. Concepts such as [ATOM] are crucial for categorizing the world.
2. Features of such concepts are theory-generated, rather than observed.
3. Definitions are revisable.
4. Conceptual change concerning atoms, molecules, and elements results from theory change.
5. The concepts of atom, molecule, and element are theoretically intertwined: they changed together as theories of matter developed.
6. Meaning is a function of both relations among concepts and indirect relations to the world via experiments.
7. Science education, and possibly also developmental psychology, must attend to the complexities of theoretical conceptual change.

We will also discuss the mental representations and processes underlying theories, explanations, and mechanisms.
2. Theories of concepts

We will first consider theories of concepts and then apply one approach to the development of atomic theory. There are five main theories of the nature of concepts in current cognitive science: classical, prototype, exemplar, neurological, and the “knowledge approach” [Medin (1989), Murphy (2002), Thagard (2005)]. According to the classical theory, a concept is defined by a set of features that are necessary and sufficient conditions of its application. For example, a shape is a triangle if and only if it has three sides. Unfortunately, such tight definitions are rare outside mathematics, and are hard to find for ordinary concepts such as [DOG] and [CHAIR], let alone for more abstract concepts such as [BEAUTY] and [JUSTICE]. Accordingly, many psychologists, philosophers, and other cognitive scientists have maintained that concepts are prototypes that specify typical features rather than necessary and sufficient conditions. Something is a dog, for example, if it matches many of the typical features of dogs such as having fur and four legs. But something can still be a dog if it lacks a typical feature. On the classical view of theories, categorization is a deductive process of reasoning with necessary and sufficient conditions, but on the prototype view categorization is an inductive process of finding the best match between the features of an object and those of the closest prototypes.

An alternative view of concepts is that they do not consist of general features but rather of stored examples. People who have observed many dogs have their observations stored in memory, and they categorize new objects as dogs based on these exemplars. The exemplar view is compatible with a neurological view of concepts as patterns of activation in neural networks. If a network stores many examples of a category, then it may have a pattern of activation that arises when similar examples are observed. However, the neural representation view of concepts is also compatible with a prototype theory, since training a network may produce patterns of activation that correspond to typical features.

The most high-level theory of concepts is the “knowledge approach,” according to which concepts are part of our general knowledge of the world, and are learned as part of our overall understanding of it. On this view, concepts are not just a matter of examples or of typical observed features, but also have a crucial explanatory role. For example, the concept of [DOG] includes features that explain why and how dogs behave as they do. The knowledge approach should be compatible with a neural representation view of concepts as patterns of activation, but requires that the patterns be achieved by processes more complex than just storage of examples or training of connections that represent typical features.

We will not consider the relative merits of the different theories for ordinary concepts such as [DOG] and [CHAIR], but it is obvious that something like the knowledge approach is crucial for theoretical concepts such as [ATOM]. The Greeks had obviously never observed any atoms; rather, they hypothesized their existence in order to explain the properties and behavior of macroscopic objects. Their mental representation of an atom included the features of being a kind of particle, of having a variety of shapes, of being in motion, and, most crucially, of being indivisible. Like atoms themselves, none
of the features of atoms are observable, so they had to be generated as explanatory hypotheses. Thus, the structure of the concept of the atom fits best with the “knowledge approach” to the nature of concepts.

The psychological purpose of the mental representation [ATOM] is to stand for a category of things, namely atoms, and to generate explanations of the behavior of all those things that are constituted by atoms. Introducing the concept of an atom was a major kind of conceptual change. The ancient alternative to the atomic theory of matter was the plenum theory, held by Aristotle and others, according to which matter is continuous and therefore can be subdivided without limit. There is no historical record of how the Greeks concocted the concept of an atom, but the most plausible cognitive mechanism is analogical abduction, a process in which puzzling facts are explained by using analogies to generate new hypotheses [Thagard (1988, pp. 60–63)]. For example, the Greeks appear to have arrived at a wave theory of sound by noticing analogies between the propagating and reflecting behavior of sound and the behavior of water waves. Similarly, we conjecture that the concept of an atom is formed analogically by noticing how the properties of big objects derive from their smaller but still observable constituents. A sword, for example, is capable of cutting because it is composed of parts, including a handle and sharp blade. In turn, these parts can be decomposed into smaller parts whose structure generates their functionality. By analogy, it should be possible to keep on subdividing down to atoms, whose fundamental immutable properties are the source of the properties and behavior of all those larger parts that they constitute. Thus, atoms are the fundamental category for understanding the world. Let us now look in more detail at the development of the concept of an atom.

3. The ancient concept of an atom

According to current estimates, the universe contains approximately $10^{80}$ atoms. We cannot directly observe the features of atoms, but physicists and chemists have come to attribute important properties to them in order to explain the structure and behavior of matter. Atoms consist of protons, neutrons, and electrons, and they bind together to form molecules. The concept of an atom has gone through at least four main stages, from the Ancient Greeks to modern quantum theory.

The story, in rough outline, proceeds from Leucippus to Linus Pauling. Leucippus, Democritus and other Greek philosophers proposed that the world consists of atoms, indivisible objects that differ in size, shape and motion. The word atom comes from the Greek words α for ‘not’ and tomos meaning ‘cut.’ On the Greek view, atoms move in a vacuum, and give substances their different properties. Atoms form substances by fitting or hooking together. Atomic theory was rejected by Aristotle, whose views were dominant for two millennia, but was revived by Gassendi in the seventeenth century and Dalton in the early nineteenth century. At the beginning of the twentieth century, Rutherford proposed that atoms are in fact divisible, consisting of both a nucleus and electrons. In 1924, de Broglie extended quantum theory to hypothesize that atoms and
electrons possess wave as well as particle properties. In the 1930s, Pauling developed a quantum theory of chemical bonds, according to which atoms are held together by the interactions of electrons as wave forms.

The ancient Greek theory of atoms was developed by Leucippus, Democritus, and Epicurus, but only fragments of their writings survive. The earliest detailed account of atoms that has survived was written by the Roman poet Lucretius (1969), who lived from about 99 to 55 B.C. According to his book, *De Rerum Natura*, atoms are the ultimate particles that constitute all things. Atoms are in constant motion and come in a great variety of shapes, which account for the different characteristics of different compound bodies. For example, olive oil flows more slowly through a filter than wine, because the atoms of the oil are larger, more hooked, and more closely intertwined than the atoms of wine. Honey and milk taste pleasant to the tongue because their atoms are smooth and round, whereas harsh and bitter substances consist of more hooked atoms that tear up the passages leading to our senses. Things that are hard and firm, such as diamonds and iron, are composed of atoms that are more hooked together than the atoms of soft substances like liquids. According to Lucretius, changes in the world are to be explained naturalistically and mechanistically in terms of the motion and shapes of atoms, not in terms of the actions of the gods.

Cantore (1969, p. 17) comprehensively summarizes the principal features of Leucippus’ and Democritus’ atoms as follows:

… matter consists of ultimate particles, which are intrinsically unchangeable and indivisible … In this view, atoms are hard, extremely small, absolutely identical corpuscles, distinct from each other only in shape and size. Macroscopic things differ from one another because of these irreducible differences among the atoms of which they are composed, and also because of the mutual arrangement of the atoms themselves. Atoms move spontaneously and ceaselessly at random in a vacuum [or void], like dust particles that can be seen dancing in a sunbeam in still air. Atoms come together out of necessity and form aggregates by a sort of hook-and-eye mechanism, not by attractive forces.

The Greek atomists also held an early principle of conservation: atoms, they thought, can neither be created nor destroyed; the ultimate constituents of the world are as fresh and un tarnished as when they were created. It is important to keep in mind that this was a metaphysician’s philosophical opinion about the nature of reality, and not an empirical hypothesis. Democritus, especially, wanted to prove that change is real, refuting Parmenides.

How does the ancient Greek concept of [ATOM] fit with current psychological theories of concepts and categories? The Greek concept of [ATOM] was clearly very different from modern theories of atomic structure, and the Greeks had nothing like our current concepts of [MOLECULE] and [ELEMENT]. The most common view was that everything consists of combinations of four elements: earth, air, fire, and water. For the atomists, these “elements” were themselves constructed out of different kinds of atoms. Below we will trace the later co-evolution of concepts of atoms, molecules, and elements. We will not attempt to recount all the changes in the concept of an atom in the long course of its evolution [Mellor(1971)], but will highlight two key developments
produced by Dalton and Rutherford. [See Asimov (1982) for details about particular scientists].

4. Revival of the concept of the atom

Lucretius’ *De Rerum Natura* was rediscovered in the fifteenth century and influenced many important thinkers, including Francis Bacon, Pierre Gassendi, and Robert Boyle. According to Mellor (1971), Gassendi was the first to use the term “molecule” to describe a cluster of atoms, but without the modern concept of a chemical element. In the eighteenth century, Lavoisier revolutionized chemistry with his development of the oxygen theory of combustion, but he did not support the atomic theory of matter, and viewed elements simply as substances that people have not yet been able to decompose. His list of elements included currently familiar ones such as oxygen, hydrogen, and iron, but also light and caloric (heat). Today, light and heat are no longer categorized as elements at all, but as processes: light is the activity of photons with wave-like properties, and heat is the result of the motion of molecules. This reclassification was a major conceptual change of the sort that Thagard (1992) called *branch jumping*, since it required a movement from one branch of the hierarchical tree of kinds to another branch.

The chemical atomic theory originated around 1800 with the ideas of the Englishman, John Dalton. From his studies of the nature of water vapor in the atmosphere, Dalton conjectured that the atmosphere consists of various gases that are mechanically mixed rather than chemically combined. In order to explain why water does not absorb every kind of gas in the same way, he hypothesized that gases differ in the relative weights of their ultimate particles (atoms). Dalton (1808) generalized that all atoms of a given element are identical and have the same invariable weight, and that atoms of different elements have different weights. The major difference between ancient concepts of atoms and Dalton’s idea was his hypothesis that the crucial property of atoms, their weight, varied consistently (in integral ratios) with different elements. The conceptual change from ancient theories was not huge, but it provided substantial new power to explain quantitative features of the behavior of gases and other substances. For Dalton, elements like oxygen were fundamental categories. In keeping with the Greek atomists, Dalton maintained that atoms are minute, discrete, indivisible, and indestructible. No means existed for determining their shape so he left questions regarding atomic structure comfortably in the arms of speculation.

Early in the nineteenth century, there was much confusion about the nature of atoms. Avogadro, for example, made no distinction between atoms and molecules. Improved understanding of atomic weights, however, enabled Mendeleev to develop the periodic table of elements. Kekulé and others developed the theory of valency that explained how atoms combine into molecules. But a major change in the concept of the atom came as a result of the discovery of the electron, and the experiments of Ernest Rutherford and his colleagues.
5. Modern development of the concept of an atom

From regularities in spectroscopic studies of hydrogen energy emission, it became clear that atoms have a complex, yet ordered structure. Decisive evidence for atomic components was first obtained from the study of cathode rays. “If an electrical discharge is sent through a highly evacuated tube (about 0.001 mm of mercury) the glass walls in the region surrounding the anode glow with a bright green fluorescent light: the agents producing the illumination were called cathode rays ...” [Cantore (1969, p. 56)].

Certain empirical features were soon discovered about these rays. Though they travel in straight lines like ordinary light, the cathode only emits them perpendicular to the emitting surface. They carry momentum, transmit energy, and are deflected by a magnetic field [Cantore (1969, p. 56)]. The most important discovery, however, was that “all of these phenomena are entirely independent of the chemical nature of the residual gas in the tube and the material constituting the cathode” [Cantore (1969, p. 57)].

What are cathode rays–material particles or electromagnetic radiation? Several studies were conducted on the electrical nature of atoms, using Faraday’s ideas on electricity and magnetism: specifically, his discovery that charges moving along magnetic lines of force are acted upon by a force perpendicular to both the current and the magnetic field. This led some to postulate that the deflection could only occur if the deflected rays were particles with negative energy.

Experiments performed by physicists on the conduction of electricity through gases led Joseph John Thompson, then director of the Cavendish Laboratory at Cambridge University, to conclude in 1897 that cathode rays are negatively charged particles with mass.

As the cathode rays carry a charge of negative electricity, are deflected by an electrostatic force as if they were negatively electrified, and are acted on by a magnetic force in just the way in which this force would act on a negatively electrified body moving along the path of these rays, I can see no escape from the conclusion that they are charges of negative electricity carried by particles of matter

Thompson (1897, p. 302).

The fundamental problem remained: Are these particles molecules, atoms, or something even smaller? By measuring the mass-to-charge ratio \((e/m)\), Thompson was surprised to find that the ratio was constant regardless of the residual gas in the tube–air, carbon, or hydrogen–or of the cathode material–aluminum, platinum, or iron [Cantore (1969, p. 58)]. At the time, it was believed that hydrogen was the smallest particle. Yet the \(e/m\) ratio for the ray was less than one thousandth that of hydrogen. And since the rays could penetrate solid matter, Thompson came to the tentative conclusion that atoms might be complex structures with smaller components; this entailed that, contrary to the views of Dalton or Democritus, atoms are not internally homogeneous or unbreakable entities. In his own words,

Thus on this view we have in the cathode rays matter in a new state, a state in which the subdivision of matter is carried very much further than in the ordinary gaseous state; a state in which all matter – that
is, matter derived from different sources such as hydrogen, oxygen, etc. – is of one and the same kind; this matter being the substance from which all chemical elements are built up.

Thompson (1897, p. 312)

Cathode rays, a new state of matter, are now known as electrons. A year before the discovery of the electron (1886), Goldstein had discovered a second kind of ray. If one pierced the negative electrode, these new rays were emitted in the opposite direction to the cathode rays. He called them canal rays. The two rays were easily distinguished. For example, if we use neon as the residual gas in the tube, a cathode ray produces a pale blue light, whereas the canal ray produces a brilliant red [Cantore (1969, p. 60)]. It was not until 20 years later, however, that Thompson (1906–1911) provided the first thorough interpretation of these rays.

If an electric and a strong magnetic field are present, the rays are deflected from their original path in a direction exactly opposite to that of cathode rays, and generally are split into several sharply distinct sunbeams. Therefore the canal rays are massive, positively charged particles …

[Cantore (1969, p. 60)]

These particles have a mass-to-charge ratio thousands of times larger than electrons. Thompson used this to speculate further on the electronic nature of the atom. Assuming the atomicity of electric charge, he argued that “the positive rays must be atoms or molecules belonging to the gas filling the discharge tube which have lost one or more electrons” [Cantore (pp. 60–61)]. This, coupled with the fact that these rays split into several distinct beams was taken as direct evidence for the discrete, well-defined values for atomic mass – this idea was in contrast to the view that atoms could occupy a continuous range of masses.

Thompson made the first proposal about atomic structure. He envisaged the atom as a “positively charged homogenous sphere with electrons distributed uniformly throughout the volume” [Cantore (1969, p. 63)] like plums embedded in plum pudding. At that time, experimenters realized that in order to probe inside the atom one requires a point-like particle satisfying two conditions: (a) it must carry a charge that interacts with the positive part of the atom, and (b) it has to penetrate the electron shield and get sufficiently close to the atom’s center. Alpha particles meet both those requirements. Alpha particles were discovered by Rutherford in 1899 and were found to contain two protons and two neutrons; they are subatomic fragments ejected from the nuclei of some unstable atoms, with a mass thousands of times larger than an electron.

In 1906, Rutherford began investigating how positively charged alpha particles are scattered by thin sheets of gold foil. His experimental setup involved an alpha particle emitter, a thin sheet of gold, and several zinc sulfide screens placed around the target to “record” particle deflection. The results of his experiment indicate that 98% of the alpha particles pass through the foil, while most of the remaining 2% deflect at wide angles, and a very small percentage, 0.01%, deflects straight back. Contrary to this result, Thompson’s model predicts that during a collision with an atom, “an alpha particle can never be deviated from its path through any large angle” [Cantore (1969, p. 65)].
In 1911, Rutherford proposed his revolutionary “planetary model.” In this representa-
tion, the atom consists of a positively charged region occupying one ten-trillionth of
the entire volume of the atom; the positively charged particles are called protons. Al-
most the entire mass, therefore, is located in the center, leaving the negatively
charged electrons with a proportionally very large space to maneuver in the outer
region. Born expressed the vastness quite eloquently:

To get an idea of the dimensions, and the emptiness, of an atom, take the following illustration. If we
imagine a drop of water to be expanded to the size of the earth, and all the atoms in it also enlarged
in the same proportion, an atom will have a diameter of a few meters. The diameter of the nucleus,
however, will be only something like 1/100 mm \[\text{Born (1957, pp. 65–66)}\].

Enormous space, coupled with the electron’s low mass, makes it easy for alpha par-
ticles to pass right through unhindered. But since atoms are ordinarily stable, a coun-
terbalancing force is needed to keep the electrons from “falling into” the nucleus due to
nuclear attraction (the “opposite charges attract” principle). Centrifugal force provides
the counterbalance if electrons are presumed to circulate the nucleus in stable elliptical
or circular orbits in much the same way that planets revolve around the sun.

Thompson’s and Rutherford’s models represented an astonishing change in the con-
cept of an atom, since from Leucippus to Dalton atoms had been taken by definition to
be indivisible. So much for definitions. In philosophical terminology, it was an analytic
a priori truth that atoms are indivisible, but Rutherford’s experiments required rejection
of indivisibility in favor of the nuclear theory. What began as an analytic truth turned
out to be false.

This development provides further support to the “knowledge approach” to the study
of concepts and categories. The concept of an atom changed dramatically, not because
of observation of atoms with novel features, but because features had to be modified in
order to generate explanations of new experimental results concerning the scattering of
radioactive particles. Once atoms had internal structure, it also became possible to
understand how different elements and molecules, composed of atoms with different
numbers of protons and electrons, could have their various properties and behaviors.

Later developments further modified the nuclear theory of the atom. In 1913, Niels
Bohr realized that orbiting electrons give off energy and should spiral into the nucleus.
To solve this problem, he combined the internal structure of the atom with Planck’s
quantum theory and proposed that electrons “jump” between energy levels, giving off
energy in the form of light when they jump to a lower level and jumping to a higher level
after the absorption of light energy. Other developments include de Broglie’s 1925
hypothesis that atoms and electrons possess wave as well as particle properties. In the
1930s, Linus Pauling developed the quantum theory of chemical bonds, according to
which atoms are held together in molecules by interactions of electrons as wave-forms.
And in 1932 James Chadwick introduced the concept of neutron. These electrically neu-
tral subatomic particles help stabilize the protons, which, when left on their own, repel
each other. And today, atoms have even become “observable” in an extended sense, in
that images of them can be produced by scanning tunneling electron microscopy.
The concepts of [ELEMENT] and [MOLECULE] have evolved along with the concept of an atom. Modern elements are defined by their atomic number – how many protons they have in the nucleus of their atoms, allowing for isotopes of the same element with different atomic weights because of different numbers of neutrons. Pauling’s quantum mechanical theory of chemical bonds explained how atoms form into molecules. Modern chemical theory provides a highly coherent explanation of a great number of observed phenomena, as well as the technological ability to predict and control many chemical reactions and processes.

6. Theories and meaning

In an early expression of the knowledge approach to the nature of concepts, Murphy and Medin (1985) considered concepts as mental theories about the world. Murphy (2002, p. 61) sensibly backed away from the term “theories,” because the beliefs of ordinary people are much less complete and consistent than scientific theories are supposed to be. However, in scientific fields such as chemistry, concepts are actually parts of theories in the full-blown sense. But what is a scientific theory?

Thagard (forthcoming) argues that most scientific theories are mental representations of mechanisms that provide explanations. The representations may be pictorial as well as verbal, so that an image of the nuclear atom as having protons surrounded by electrons is as much a part of the nuclear theory as a verbal description. Mechanisms are systems of parts related to each other in ways that produce regular changes. Explanations of an event consist of describing a mechanism in such a way that the event is produced by the interactions of the parts of the mechanism. Instead of relying on theological and teleological explanations, the ancient Greeks brilliantly perceived an analogy between human-constructed machines such as levers and natural phenomena such as the motion of objects. All of the explanations offered by the various stages of atomic theories are mechanistic in this way.

Concepts like [ATOM] acquire their meaning from their places in such theories. Atomic theory gives meaning to the concept of an atom by relating it to other concepts, such as [SHAPE], [MOTION], [DIVISIBILITY], and (more recently) [PROTON] and [ELECTRON]. Ideally, however, concepts are supposed to get meaning not just from their relation to other concepts, but also from their relation to the world: they should have reference as well as sense. It is hard to say whether the Greek concept of an atom referred to anything in the world. On the one hand, we now know that there are no variously shaped, hooked, indivisible particles of the sort that the Greeks envisioned. On the other, objects really do consist of basic particles that are responsible for their macroscopic properties. So something like the Greek atom does exist. The situation is less ambiguous with the modern concept of atom, which derives its meaning both from its place in contemporary chemical and physical theories and from its empirically substantiated relation to the world. It was only with the experiments of Thompson and Rutherford that scientists entered into a kind of direct interaction with individual atoms, by virtue of their ability to shoot rays that penetrated them.
So what are concepts and how can they represent categories? For theoretical concepts such as [ATOM], we suggest that they be viewed as abductive prototypes. As we described above, prototypes differ from classical definitions in that they specify typical features rather than necessary and sufficient conditions. Abductive inference, or abduction as it was dubbed by the philosopher C.S. Peirce, is inference in which a hypothesis is generated in order to explain something puzzling. The features of abductive prototypes are hypothesized in order to explain observations, as when Rutherford inferred that the mass of an atom is concentrated in a very small region in order to explain why alpha particles pass through gold foil. Abductive prototypes can change dramatically when new data require revision of hypotheses concerning explanatory features. This is just what happened to the concept of an atom when the experiments of Thompson and Rutherford revealed the divisibility of atoms.

7. Conclusion

When Thagard’s son Adam was about four years old, he was told that he was made up of tiny atoms. He misheard the term, and promptly conjectured that his brother Daniel was made up of little daniels. Researchers in science education have documented more serious problems that children and older students have in acquiring the modern scientific concept of an atom [Griffiths and Preston (1992), Harrison and Treagust (2001)]. Given the more than two millennia that it took for the modern concept to evolve, it is not surprising that students do not automatically acquire important theoretical concepts like atom, molecule, and element.

Nevertheless, acquisition of such concepts is an essential part of achieving the ability to explain the world and to divide its constituents into empirically plausible categories. We have described how changes in the concepts of [ATOM], [MOLECULE], and [ELEMENT] have been an important part of the development of theoretical categorizations of the chemical and physical world. Such concepts are not strictly defined, nor are they prototypes of observed properties or sets of exemplars. Rather, they have features that are postulated in order to foster the explanatory role that is central to theoretical concepts, which are used to characterize the mechanisms that provide explanations. Thus, progress in categorization depends on theoretical conceptual change.

References


Chapter 11

RELATIONS BETWEEN LANGUAGE AND THOUGHT: INDIVIDUATION AND THE COUNT/MASS DISTINCTION*

ANNA PAPAFRAGOU
Department of Psychology, University of Delaware

Contents

Abstract 256
1. Introduction 257
2. Strong discontinuity proposals 260
   2.1. Quine 260
   2.2. Abstract individuation in language and thought 262
3. Weak discontinuity proposals 263
   3.1. Crosslinguistic studies 263
   3.2. Language-on-language effects 266
4. Material and shape cues in labeling and categorization 268
5. Conclusion 271
References 271

* I would like to thank Peggy Li for very useful discussions of this material. Writing of this chapter was partly supported by NIH/NRSA Grant #F32MH065020.
Abstract

What is the relationship between linguistic and nonlinguistic cognitive categories? How does language acquisition (specifically, the acquisition of grammatical categories) draw on prelinguistic concepts? Is it possible, as recent commentators have argued, that the acquisition of linguistic categories itself affects nonlinguistic conceptual categories? This chapter addresses these questions by focusing on the grammatical distinction between count and mass nouns and its relation to the distinction between objects and stuff (materials). We first ask whether learning count/mass syntax may help children think about objects and stuff in ways that were not antecedently available to them. We also ask whether crosslinguistic differences in marking count/mass status affect the nonlinguistic individuation criteria used by speakers of different languages. We review a number of recent findings that have been interpreted as showing such effects of count/mass syntax on nonlinguistic cognition, and we argue that they do not conclusively demonstrate language-specific influences on mental life.
1. Introduction

Several researchers within the fields of linguistics and developmental psychology hold that human concepts and mental architecture do not change in crucial respects throughout human development – in other words, the mental representations of children do not differ in fundamental ways from those of adults [Fodor (1975), Macnamara (1982), Pinker (1984), Gleitman (1990)]. These *continuity* theorists are impressed by the fact that prelinguistic infants already possess a rich inventory of conceptual categories – that are presumably part of the universal human mental apparatus – including notions of space [Levine and Carey (1982), Needham and Baillargeon (1993)], event cognition [Gergely and Csibra (2003), Woodward (2004)], quantity/number [Gallistel and Gelman (1992)], causality [Bullock and Gelman (1979), Leslie and Keeble (1987)], agency and animacy [Gelman and Spelke (1981), Woodward, Phillips and Spelke (1993)], among many other notions. Prelinguistic primitive notions such as these are assumed to form the basis for the acquisition of grammatical categories. As Chomsky (1982) has forcefully argued, without the existence of such primitive concepts, it is hard to see how language acquisition would take place at all:

> The claim that we’re making about primitive notions is that if data were presented in such a way that these primitives couldn’t be applied to it directly, prelinguistically, before you have a grammar, then language couldn’t be learnt... We have to assume that there are some prelinguistic notions that can pick out pieces of the world, say, elements of this meaning and this sound.  
*Chomsky, 1982 (unknown)*

On this view, language acquisition essentially presents children with a mapping problem: the main task of the learner is to figure out which aspects of the input language correspond to which nonlinguistic conceptual primitive notions – or combinations thereof. Specific proposals have addressed the question of how the mapping problem may be solved at the very early stages of language development [see Pinker (1984), Gleitman (1990)].

Other researchers argue that the human conceptual typology can and does undergo change during development. On this view, cognitive development is characterized by deep *discontinuities* in mental architecture. There are several accounts of the exact nature of such discontinuities, and of the way they affect specific conceptual domains [for explanations of how such discontinuities affect our notions of space, number, or mental representation, see Spelke and Tsivkin (2001), Carey (2001), and Gopnik and Wellman (1996), respectively.] Furthermore, a growing class of discontinuity theories considers language itself (in particular, the acquisition of grammatical categories) to be a major cause of conceptual change. In the words of Bowerman and Levinson:

> Instead of language merely reflecting the cognitive development which permits and constrains its acquisition, language is being thought of as potentially catalytic and transformative of cognition.  
*Bowerman and Levinson, 2001, pp. 12–13*
On this approach, language learning does not simply depend on prior representational resources, but it carries the potential of affecting those resources by introducing or otherwise modifying an individual’s concepts. Carried to its logical conclusion, this view predicts that there should be deep conceptual differences, not only between young and experienced learners of a language (that is, between children and adults), but also between speakers of different languages. In this sense, recent proposals about language-driven discontinuities bear a more or less close relationship to the theories of B. L. Whorf (1956), who famously argued that the grammatical categories of a language affect (or even streamline) the conceptual life of its speakers. There is currently a growing body of theoretical and experimental work which looks for linguistic influences on conceptual organization and cognitive development from a neo-Whorfian perspective [Gumperz and Levinson (1996), Levinson (1996a,b), Bowerman and Levinson (2001), Gopnik (2001), Gentner and Goldin-Meadow (2004); for a critical discussion, see Gennari, Sloman, Malt, and Fitch (2002), Li and Gleitman (2002), Munnich, Landau and Dosher (2001), and Papafragou, Massey and Gleitman (2002), among many others].

In this chapter, we want to contribute to recent discussions of the language-thought relationship by investigating how language, and especially the acquisition of grammatical categories, relates to conceptual organization. We focus on a specific case, the grammatical distinction between count and mass expressions. “Count” expressions refer to individual objects (e.g., We need these tables); or, they refer to kinds of individual objects (e.g., Tables are furniture). “Mass” expressions, on the other hand, refer to portions of quantities (e.g., There is a lot of water on the floor); or, they refer to kinds of quantities (e.g., Water is found in the sea). Put differently, count nouns refer to discrete, well-delineated groups or entities, while mass nouns do not make explicit how their referents are to be divided into objects. Syntactically, count expressions co-occur with quantifiers such as each, every, many, several, few, (stressed) some and the indefinite article a(n); they use counting phrases (e.g., five, a score of); and they can be pluralized. Mass expressions, on the other hand, co-occur with quantifiers such as little, much, and (unstressed) some; they use measurement phrases (e.g., liters of); and they cannot be pluralized.

This grammatical distinction is linked to the conceptual distinction between objects and substances: crosslinguistically, objects tend to be named by count nouns (e.g., cat, table), and substances by mass nouns [e.g., milk, wool; see Markman (1985)]. However, objects and materials do not exhaust the range of referents for count and mass nominals respectively. In English, for instance, count nouns name objects (e.g., a dog), but also abstract entities (e.g., an indication), or events (e.g., an adventure). Similarly, English mass nouns denote substances (e.g., water), but also abstract entities (e.g., beauty), solid materials (e.g., wood), or unindividuated groups of objects (e.g., jewelry).

For these reasons, the semantic underpinnings of the count/mass distinction are properly characterized in terms broader than the object/substance distinction [as a number of commentators have noticed; see Chierchia (1998), Gillon (1992), McCawley (1975) and Pelletier and Schubert (1989)]. In a widely held view, count/mass syntax maps onto a quantificational distinction between individuals and nonindividuated entities [Link
The cognitive notion of individual relates to properties such as countability, indivisibility, and boundedness, and it corresponds approximately to ‘discrete bounded entity.’ Possible individuals may be material objects (e.g., a dog), events which take a bounded interval of time (e.g., a race), mental states (e.g., a migraine) or temporal stretches (e.g., a week); see Bloom (1994a) for an extensive discussion. This way of semantically characterizing the count/mass distinction has the advantage of capturing the denotation of the entire class of mass and count nouns, as discussed above. Moreover, it makes linguistic sense. From a formal quantificational point of view, there is no difference between moment and mouse, or between seriousness and snow: the first pair, unlike the second, contains nouns that denote kinds of individuals and can form more complex nominal phrases (NPs) to denote single individuals (e.g., a moment, a mouse).

The precise form of the link between linguistic individuation (count/mass marking) and nonlinguistic ontological categories has given rise to intense debate within linguistics and philosophy, and we will review some of the relevant arguments in the sections to follow. More crucially, for present purposes, this link has become fertile ground for several discontinuity proposals. These proposals have come in two main varieties. Strong discontinuity theorists have proposed that the quantificational system of natural language (including the count/mass distinction) helps children arrive at the ontological distinction between individuals and nonindividuals. Other, weaker discontinuity proposals are consistent with the existence of a universal (object/substance) ontology, but they explore the possibility that typological differences among languages in encoding count/mass status may affect the salience, or the boundaries, of our ontological categories.

Our goal in what follows is to reconsider the arguments that have been used to support such discontinuity theories. We begin by asking whether learning the count/mass distinction may help children think about objects and materials in ways that were not antecedently available to them. We present several lines of experimental results that give us reasons to think that the count/mass distinction presupposes, rather than introduces, basic ontological distinctions. In the second part of this chapter, we discuss whether the weaker view may be true – that is, whether crosslinguistic differences in count/mass categories affect the individuation criteria of speakers of different languages. We review a number of recent findings that have been interpreted as showing such effects of count/mass syntax on nonlinguistic cognition, and we argue that they do not conclusively demonstrate language-specific influences on mental life.

1 We should point out that there are several different proposals about the semantics of mass nouns, not all of which share the idea that mass denotation excludes individuation; see, for example, Barner and Snedeker (in press), Chierchia (1998), and Gillon (1992).

2 This view also seems to account for certain empirical facts about how adults spontaneously interpret count and mass syntax. In one experiment, adults were taught novel words referring to sensations or sounds [Bloom (1994b)]. The syntax of the word was kept neutral between count and mass status. The new word was offered as a description of either something that occurs in discrete units of time (temporal individuals) or of something that occurs over continuous periods of time (unbounded, unindividuated entities). As predicted, subjects categorized words for temporal individuals as count nouns and names for temporal stretches as mass nouns.
2. **Strong discontinuity proposals**

2.1. *Quine*

Perhaps the most radical proposal about how a grammatical category such as the count/mass distinction may affect the development of human conceptual categories comes from Quine (1960). According to Quine, the ontological distinction between objects and substances that underlies language is a cultural construction. In other words, before mastering the relevant aspects of their language, children do not represent the world in terms of stable objects, but rather as histories of sporadic encounters, encompassing an undifferentiated portion of what goes on. For instance, prelinguistic infants do not conceptualize a dog as an individuated, countable entity, but as an instance of doghood. They may still interact with it, learn several things associated with it (e.g., that it barks), and use shape and other criteria to identify it. Still, in order to isolate and individuate a dog, infants need to use the resources made available to them through the quantificational system of natural language: by being exposed to phrases such as *a dog, these dogs*, etc. – that carve out boundaries of perceived experience, the child could find a way of bootstrapping into the adult ontological system [see Carey (1994) for discussion]. According to Quine, then, the count/mass syntax offers the means for discriminating objects versus substances during development.³

Soja, Carey and Spelke (1991) designed an experiment to address Quine’s claim that knowledge of individuation is a product of mastering the count/mass distinction. They taught English-speaking 2-year-olds who did not show productive command of the count/mass syntax, novel words in reference to various stimuli. In each case, the target stimulus was named by a nonsense word embedded in a frame that did not mark it for count/mass status (e.g., *This is my blicket.*). Under one set of conditions, the stimulus was a solid object (e.g., a pyramid made of wood). The children were then shown two alternatives: one alternative had the same shape as the original, but was made out of a different material (e.g., a pyramid made out of sculpting material); the other alternative had a different shape but was made of the same material (e.g., pieces of wood). When asked to choose which of the alternatives was the *blicket*, children in this condition consistently chose the same-shape alternative.

Under the second set of conditions, the target stimulus was some nonsolid substance (e.g., face cream) arranged in an interesting shape. Again, children were shown two alternatives, one that maintained the shape of the target but consisted of a different material (e.g., hair-setting gel), and another which consisted of piles of the target substance (e.g., piles of face cream). Children generalized the target name to the same-substance display. On the basis of these results, Soja et al. (1991) concluded that children

³ A similar view is found in the writings of B.L. Whorf, who argued that the individuated/nonindividuated distinction “is somewhat forced upon our description of events by an unavoidable pattern in the English language [i.e., the count/mass syntax]” [Whorf (1956), p. 141].
possess innate ontological commitments that guide the extension of novel words for objects and substances, prior to the acquisition of count/mass syntax.

There is further compelling evidence that, pace Quine, the logical resources to represent objects are available very early in life. Several studies have shown that even young babies use spatiotemporal criteria for individuating solid, small, moveable, coherent objects. In one experiment [Spelke, Kestenbaum, Simons and Wein (1995)], babies of 4½ months were first shown two screens side by side. Then an object appeared from the left edge of the left screen and went back behind it; after a suitable interval, a second object appeared from the right edge of the right screen and returned behind it. When the screens disappeared, babies looked longer at (i.e., were surprised by) displays featuring one object rather than two. This shows that babies, just like adults, reason that objects cannot be in two places at the same time, and that objects continue to exist behind opaque screens [Mehler and Fox (1985), Spelke (1985, 1990), Xu and Carey (1996)]. Other work has shown that babies at this age can distinguish one object from two numerically distinct but physically similar objects [Wynn (1992, 1995), Baillargeon (1993), Uller et al. (1999)]. This research shows that the concept OBJECT is not constructed through experience with natural language quantification, but is probably part of the core knowledge of the human conceptual system.

More recent evidence comes from the work of Susan Carey and her colleagues, who have shown that very young infants possess something close to an object/substance distinction. In one of these studies [Carey (1994, 2001), see also Huntley-Fenner (1995), Huntley-Fenner, Carey and Solimando (2002)], 8-month-old babies were familiarized with either a solid object in the form of a sand object, which was suspended by a narrow thread and moved around as a coherent whole (the Object condition) or with a quantity of sand (the Sand condition). In the Object condition, babies saw a sand object being put behind a screen, then a second sand object lowered behind the screen. When the screen was removed, babies showed no surprise when two objects appeared, while they showed surprise at the unexpected outcome of one object (which occurred when one object was taken away). In the Sand condition, babies saw an amount of sand being poured onto a surface and then hidden behind a screen. More sand was poured either on top of or beside the original pile of sand behind the screen, and then the screen was lifted. Babies showed no surprise at the unexpected outcome of one pile of sand but looked slightly longer at the expected outcome of two piles of sand, which was also the baseline preference. It is reasonable to assume that the reason babies fail the test in the Sand condition is the noncohesiveness of the sand. If so, these experiments show that preverbal infants draw a spontaneous and systematic distinction between individuated and nonindividuated material objects. Hence these babies appear to possess some rudimentary form of a distinction, which they may presumably use in order to acquire the count/mass distinction in the English language.

We conclude that infants’ individuation capabilities overall support the continuity/universality approach to grammaticizable concepts: pace Quine, early individuation capacities predate the acquisition of the quantificational count/mass distinction in natural language.
2.2. Abstract individuation in language and thought

Recall that the count/mass distinction corresponds to a much more abstract individuation distinction than the object/substance distinction. One question worth raising at this point is whether young children are sensitive to these more abstract individuation considerations. It could be hypothesized that children start out attending to – and learning names for – highly individuatable concrete referents, and only later come to grasp the notion of an abstract individual. There are two lines of findings which speak against this idea: the first concerns the individuation capacities of prelinguistic infants, which seem to extend beyond a simple object/substance distinction; the second involves young children's ability to map these complex individuation concepts onto the count/mass syntax.

There are good reasons to believe that the appreciation of abstract individuated entities is not a late cognitive achievement. As we hinted earlier (see Section 2.1), infants have been shown to individuate a variety of nonobject, nonphysical entities. The main evidence comes from infants’ ability to count. Counting is usually taken as evidence for individuation, since in order to count something, one has to know what it is they are counting (or at least discriminate among distinct occurrences of the counted entity). Infants can count very diverse entities, including randomly arranged household objects depicted in photographs [Starkey, Spelke and Gelman (1990)], sounds [Starkey et al. (1990)], and actions such as jumps [(Wynn 1996)]. If this is true, then prelinguistic children must possess a complex notion of individuated entity, which may be comparable to the sophisticated conception that has been argued to be part of adult cognition.

Of course, nobody would deny that objects and substances are prime examples of individuated and nonindividuated entities, respectively. It is very plausible that we are cognitively biased to treat discrete objects as canonical individuated entities. Young children are particularly prone to this bias: Shipley and Shepperson (1990) found that when children are presented with an array of objects and asked to count colors or parts, they go on to count objects. In the context of word learning, it has been shown that children have trouble learning words for solid substances (e.g., wood), a fact that is explained by the presence of the cognitive bias to treat objects as canonical individuated entities [Prasada (1993), Bloom (1994b)]. Notice, however, that this bias can be overridden: as we saw, young children can count not only objects but also noncanonical individuated entities such as sounds and actions.

There is also evidence that the count/mass syntax in young children does not build directly on the object/substance classes, but instead connects to the richer individuation system outlined above. In fact, object names make up a higher proportion of the vocabulary of children than that of adults [Gentner and Boroditsky (2001), Macnamara (1982)]. But children’s early count nouns already refer to much more besides objects: there is considerable evidence that 20-month-olds use words such as bath, breakfast, friend, day or uncle [Nelson, Hampson and Shaw (1993)].

More importantly, children can use syntactic marking as a cue to individuation. In a classical study, Brown (1957) showed preschoolers a picture of an unfamiliar action performed with an unfamiliar object on an unfamiliar substance. One group of children
was told: “In this picture, you can see a sib. Now show me another picture of a sib.” Another group was told: “In this picture, you can see some sib. Now show me another picture of some sib.” When count syntax was introduced, children understood the nonsense noun to refer to the object, but when mass syntax was present, children interpreted the noun as referring to the substance. Such syntactic cues have been shown to be very powerful. When intuitions about individuation contrast with the count/mass status of a novel noun (e.g., when an unfamiliar substance is named by a count noun), very young learners override perceptual/cognitive factors in favor of the syntactic cues [Gordon (1985, 1988), Subrahmanyam, Landau and Gelman (1999)].

More strikingly, perhaps, children can use count/mass syntax to infer the name of kinds of individuated entities that are not objects. Soja, Carey and Spelke (1992) studied 2-year-olds who had a productive command of count/mass syntax. She found that they can use a word’s mass status to infer that the word refers to a nonsolid substance (e.g., water) or use count status to infer that its referent is a bounded entity made of that substance (e.g., puddle). Bloom (1994b) found that 3-year-olds will interpret a plural count noun that describes a series of actions (e.g., “These are feps”) as referring to the individual actions, and will interpret a mass noun describing the same series (e.g., “This is fep”) as referring to the undifferentiated activity. Finally, Bloom and Kelemen (1995) showed that adults and older children can use count-noun syntax to learn names for novel collections, terms such as flock and family that refer to groups of distinct physical entities.

To summarize, the evidence just discussed suggests that children may possess individuation criteria that approximate those of adults, and that these criteria (rather than a cruder object/substance distinction) form the basis for their semantic interpretation of the count/mass distinction. These data offer further support against strong discontinuity proposals which claim that early individuation criteria are provided by the linguistic distinction between count and mass nouns.

3. Weak discontinuity proposals

Several recent studies have pursued the question whether crosslinguistic differences in marking the count/mass distinction affect speakers’ individuation patterns. Overall, these studies recognize that there may be strong constraints on children’s initial presuppositions about individuation, but they leave open the possibility that language may also play a causal (albeit weaker, more restricted) role in determining individuation criteria.

3.1. Crosslinguistic studies

One series of studies comes from John Lucy and his colleagues [Lucy (1992, 1996), Lucy and Gaskins (2001)], who have conducted comparative research with individuals who speak English and those who speak Yucatec, Mayan language a Yucatec (like Chinese, or Japanese) is a classifier language: its nouns are all mass. In Lucy and Gaskins’ words, they are ‘semantically unspecified as to quantificational unit’
The language only optionally marks plural on certain nouns – mainly those denoting animate entities; for most nouns, there is no alternation of the kind ‘table/tables.’ Moreover, numerals are not able to combine directly with nouns: a classifier is necessary to individuate units which may then be counted. For instance, in this language, one cannot say ‘one banana,’ but rather has to say ‘one portion of banana’ (cf., un-tz’iit kib’– ‘one long, thin candle’ vs. ka’a-tz’iit kib’– ‘two long, thin candle’).

On the basis of these grammatical properties, Lucy and his team hypothesized that Yucatec and English nouns draw the speakers’ attention to different properties of the nouns’ referents. Their reasoning goes as follows. In order to refer to any discrete object\(^4\), English speakers use mostly shape considerations to determine which count noun is most appropriate to name the object (e.g., candle). But speakers of Yucatec, since nouns in their language do not typically denote individuated objects, cannot use shape as a criterion for naming objects; instead, they must take into account the objects’ material composition (e.g., kib’ – ‘wax/candle’). Given that these criteria for verbally labeling objects are repeatedly and heavily used in the two speech communities, it is plausible that members of these communities develop distinct preferences for attending to shape versus material, even in nonlinguistic tasks involving individual objects. Moreover, it is likely that adults will be more “contaminated” by these language-specific preferences than younger learners of English or Yucatec, since the adults have had much longer exposure to the naming practices of the community.

These predictions were tested in a number of experiments. In a typical experiment, adult participants were presented with a triad of everyday objects consisting of an original pivot object (e.g., a small cardboard box) and two alternate objects—one that had the same shape as the pivot object (e.g., a small plastic box) and another that had the same material as the pivot object (e.g., a small piece of cardboard). When asked to choose which of the alternate objects the pivot object was more similar to, Yucatec-speaking adults were more likely than English-speaking adults to choose the same-material alternate, while the English-speaking adults were more likely to pick the same-shape alternate. The triads experiment was also used to test the developmental predictions outlined above. It turned out that, while both very young English-speaking and Yucatec-speaking children preferred shape-based classifications, by the age of 9 children in the two communities had converged on the adult classification patterns: English-speaking children made shape-based choices and Yucatec children made material-based choices.

A related line of findings comes from word learning studies asking whether crosslinguistic differences in learning nominals reveal underlying differences in ontological assumptions. Imai and Gentner (1997) used as a starting point Soja et al.’s 1991 and 1992 results. The authors reasoned that if indeed the object/substance distinction is made prior to and independently of count/mass syntax, then individuals who learn classifier languages like Japanese should behave just like those who learn English when extending

\(^4\) Reference to nondiscrete objects (substances) is supposed to be more similar among these two languages, so no major cognitive differences are expected.
words to novel referents. Imai and Gentner tested monolingual Japanese and American children and adults using Soja et al.’s tasks. They found that American and Japanese adults and children as young as 2 years extended names for complex objects (e.g., a porcelain lemon juicer) on the basis of shape rather than substance. This finding replicates Soja et al.’s results and offers strong support for early ontological categories (and against Quine’s strong discontinuity conjecture). But Imai and Gentner also found that, in naming simple objects (e.g., a cork pyramid), speakers of the two languages diverged: English-speaking children from age 2 onward predominantly named objects by paying attention to shape, while Japanese children were at chance. Differences were also found in naming substances (e.g., sawdust): only the Japanese subjects extended the label on the basis of material, while the American subjects were at chance. Based on these results, Imai and Gentner concluded that universal ontological categories guide word learning for canonical objects and materials; however, categorization for items that fall in between these two canonical kinds (i.e., simple objects which could in principle be conceptualized either as substances or as individuated objects) is determined by factors specific to the language being learned [cf. also Gentner and Boroditsky (2001)]

Taken together, these crosslinguistic experiments have been interpreted as evidence for the position that, alongside universally shared ontological commitments (broadly, a distinction between individuated versus nonindividuated entities), there are language-specific effects on the boundaries of nonlinguistic ontological categories (specifically affecting the classification of simple objects). Furthermore, such effects of language are claimed to increase with age and exposure to the language. The precise interpretation of these results is, however, not straightforward. Other research shows that simple and complex objects are not treated differently by our prelinguistic individuation mechanisms. For instance, infants have no difficulty individuating simple objects made of sand alongside complex (i.e., structured and purposeful) objects and substances [Huntley-Fenner (1995)]. This result is hard to reconcile with the claim that the individuation of simple objects is indeterminate and therefore open to linguistic influences.

There is also considerable debate surrounding the proper treatment of the count/mass distinction (see Section 2.2). Several authors have pointed out that the distinction is not absent in classifier languages, but it shows up in places other than noun reference [e.g., in the kind of classifier used; Cheng and Sybesma (1999)]. More crucially, as we saw in the Introduction, many commentators have emphasized that the denotation of mass nouns is abstract and cannot be directly linked to notions like “substance” [McCawley (1975), Pelletier and Schubert (1989), Krifka (1995), Chierchia (1998)]. Thus it is unlikely that the denotation of nouns in classifier languages can give rise to a “material bias.”

Some aspects of the word learning results are more difficult to explain. For instance, since substances are claimed to be a canonical category in the universal ontology (rather than a case of indeterminate individuation) the difference between American and Japanese subjects in naming substances is unexpected. Note incidentally that these findings differ from what Soja et al. found for American subjects.
This last point has been confirmed by studies examining the interpretation of nouns in classifier languages. In a recent study by Imai and Gentner (2003), Japanese children were first introduced to and given a label for an item (e.g., a wooden pyramid). Then they were given choices that were either the same in shape (e.g., a cork pyramid), or in material (e.g., wooden pieces), or were distractors (e.g., a wax kidney bean). The children were given free reign to choose as many matches as they wanted, but they preferred only one choice. If they picked the shape match, they did not pick the material match, and vice versa. This result is in contrast to another condition in which the item was introduced, but never labeled. In this latter condition, children often picked multiple matches (i.e., both the shape match and the material match). Thus, Japanese children think that nouns either name a kind of object or a kind of substance, but not both. This importantly debunks the notion that nouns have to refer to substance kinds in a classifier language. These data also show that young speakers of classifier languages perform differently in linguistic and nonlinguistic categorization tasks, a point we return to in the next section.

Further studies have compared children’s earliest vocabularies in English and Japanese. Colunga and Smith (2000) asked adult native speakers to determine whether the nouns young children know refer to entities that had similar shape or similar material, and whether they refer to solids or nonsolids. Of interest is whether compared to English, more nouns in Japanese refer to material. Colunga and Smith found similarities in the types of noun repertoires in these two languages. The majority of the nouns refer to solid entities. Furthermore, nouns in both languages more often refer to categories high in shape similarity and less often to categories high in material or color similarity. Overall, Japanese and English share roughly equal proportions of “shape-based” and “material-based” nouns. Similar results have been obtained by looking at maternal input in English and Mandarin [Sandhofer, Smith and Luo (2000)]. Collectively, the Japanese, Mandarin, and English data support the conclusion that the nouns children learn and those they hear in their environment generally do not refer to items of similar materials.

3.2. Language-on-language effects

Given the results described above, how are we to interpret the data from crosslinguistic studies on individuation? In the next few paragraphs, Fisher and Gleitman (following) and Gleitman and Papafragou (2005), we consider an alternative explanation of these findings — one that does not rest on linguistic influences on thought.

Beginning with the word-learning studies, recall that Imai and Gentner’s task involved making guesses about the referents of novel nouns. It is reasonable to assume that such guesses must have been affected, among other things, by English and Japanese speakers’ knowledge of how linguistic forms map onto meanings in their native language. We know from prior work that people’s guesses about the referents of novel verbs [Naigles and Terrazas (1998)] or adjectives [Waxman, Senghas and Benveniste (1997)] is affected by knowledge of language-specific syntax-semantic mappings. We also know that children are sensitive to probabilistic expectations about
what the words in their language mean, or how they are combined into sentences [Choi and Bowerman (1991), Slobin (1996, 2001)]. Such probabilistic expectations guide how novel words will be interpreted, especially in contexts that supply poor or ambiguous information as to what the novel word could name.

In light of these observations, let us reconsider the word-learning findings. As we saw, subjects in both languages picked the “object” interpretation for a novel noun when the materials included nonaccidental-looking structure-rich objects. However, when the display included uninformative, “simple” objects, which could in principle be considered either as objects or as substances, subjects fell back upon language information to solve the task. Japanese does not formally distinguish between object or substance expressions, so Japanese speakers chose at random for those ambiguous stimuli. For English-speaking individuals, the presentation conditions sought to replicate this indeterminacy by avoiding count- or mass-biasing syntax (e.g., *This is my blicket*).

However, given the much higher proportion of count nouns in English, it is likely that American participants construed the novel label as a new count noun – referring to the kind of object displayed – rather than a new mass noun. If this is true, then the English-Japanese difference in novel word extensions is not an effect of language on thought, but rather an effect of “language on language” [See Fisher and Gleitman (2002) and Gleitman and Papafragou (2005) for discussion.]

It is an open possibility that such linguistic considerations intrude into otherwise nonlinguistic memory/categorization tasks and affect subjects’ individuation preferences. Imai and Mazuka [1997; reported in Imai (2000)], using a nonlinguistic version of the Imai and Gentner tasks, found very similar results to the original word extension experiments: in picking the object that was “the same” as the pivot, English adults focused on common shape and Japanese adults focused on common material. As we saw, Lucy and his colleagues report similar results for nonlinguistic tests comparing Yucatec- versus English-speaking individuals. A plausible hypothesis is that, in the absence of clear grounds on which to perform classification (since what counts as “the same” may vary indefinitely), adults implicitly use verbal mediation to solve these tasks – thereby reproducing the results of the linguistic (word extension) studies. 

6 Cultural considerations may also bias speakers for or against shape- or material-based classification. The English-speaking population in Lucy’s tasks lives in a culture where shape is very important, from children’s toys to traffic signs. By contrast, Yucatec Mayans are part of a rural community where the material of many everyday objects is important and salient. This difference might explain why, in more recent versions of the triad tasks (which involved sorting), Lucy and Gaskins find clearer preference for material on the part of the Mayan adults. According to the authors, in these new tasks, 

[the Yucatec speakers were constantly evaluating the material composition of the test items before sorting them: feeling how heavy they were, poking their nails into them to test for malleability, scraping the surface to see what the material under the paint was, smelling and tasting the objects… —and all this with familiar objects.

Lucy and Gaskins (2001, p. 272)

Similar studies that have removed the cultural differences from the subject populations (by studying urban Chinese and English speakers) have failed to replicate Lucy’s results [Mazuka and Friedman (2000)].
Support for this hypothesis comes from the fact that, in both Lucy’s and Imai and Mazuka’s nonlinguistic experiments, adults make language-consistent categorization choices, but children do not. For instance, recall that in Lucy’s studies, only by the age of 9 did Yucatec speakers adopt the adult-like preference for material in categorization tasks. This difference is unexpected on the view that linguistic influences of the count/mass syntax are responsible for adult categorization patterns. After all, children have mastered the basics of the syntax of classifier languages well before the age of 9. The findings, however, make sense for the alternative hypothesis, since children, unlike adults, may be unable to fall back on linguistic labels to solve nonlinguistic (memorial or categorization) tasks. This is in line with memory research showing that young children do not spontaneously co-opt linguistic representations in support of memorial recall, and that this ability undergoes considerable developmental changes [Hitch, Woodin and Baker (1989), Palmer (2000)].

This line of argument predicts that speakers of classifier and nonclassifier languages should behave identically in nonlinguistic individuation tasks that test for intuitions about kind membership without evoking linguistic mediation. In a recent study, Li and Dunham (2005) asked English, Japanese, and Mandarin speakers to use a scale of 1 to 7 to rate how likely they were to construe a novel specimen as a kind of object or a kind of substance. This method avoids vague instructions and is overall less open to linguistic intrusions. Results show that adults in all three languages gave similar ratings. (Interestingly, when the same experimental stimuli were embedded as targets in a standard match-to-target triad task, the language difference reappeared, with English speakers making more shape-based choices.) These results offer compelling evidence that cognitive individuation criteria are independent of language-internal distinctions, such as those encoded in count/mass syntax.

4. Material and shape cues in labeling and categorization

Much of the literature on the count/mass distinction (and on word learning, more generally) we have reviewed often makes the following two assumptions. First, at least for some classes of count nouns (e.g., those that refer to discrete objects), shape is typically the basis for extending the noun to other members of the referent class (what has been called the shape bias). Second, at least for some classes of mass nouns, material is typically the basis for extending the noun to other members of the referent class (what can be called the material bias). Before we conclude, we would like to discuss the role of shape and material considerations in naming and categorization. Where do these “biases” come from? Are they language-based generalizations? On weak discontinuity views, at least, children and adults who speak languages with count nouns come to map perceptual properties of objects (such as shape) onto count syntax as a result of their experience with count nouns, and hence develop a preference for extending count nouns to objects of similar perceptual contour. Similarly, children and adults who speak classifier languages with exclusively mass nouns come to map the material constitution of objects (i.e., the material) onto mass syntax as a
result of their experience with mass nouns, and hence are more likely to extend novel nouns to objects of similar material composition. The question we want to address now is whether the shape and the material bias are, in fact, tied to word learning, or, more specifically, to the acquisition of the count/mass distinction.

Beginning with mass nouns, notice that the “material bias,” if true, has limited applicability: it only holds for mass nouns with physical referents and not for abstract nouns such as information or nobility. Furthermore, there are exceptions to this bias even within the class of concrete, physical mass nouns. Perhaps the largest class of relevant counterexamples in English comes from mass superordinates (food, jewelry, footwear, etc.): the decision to call a plastic chair, a wooden table and a leather sofa furniture is explained by the fact that these entities all share characteristics that relate to their creation, purpose, and everyday use. The general conclusion then is that, although “concrete” in reference, superordinate mass nouns refer to objects that share not material constitution but rather functional properties – and this seems to be true crosslinguistically [Markman (1985), Wisniewski et al. (1996)].

Even in cases of mass nouns with basic-level, observable referents, the decision to name an object with a mass noun takes into account far more than simple material constitution. Suppose you find out that what you considered to be tea in your cup was in fact water from the tap that had passed through a tea filter at the reservoir. It is unlikely that you would go on calling it tea, even though its contents may be identical to the ingredients of an ordinary cup of tea [Chomsky (1995)]. What determines labeling here, as in so many other cases, is sameness of kind (rather than sameness of material). Our reluctance to call the new substance tea comes from our reluctance to classify it as an instance of that kind of beverage. What will be classified as an instance of a kind is often not easy to determine (and may vary depending on one’s perspective, goals, etc.). The lesson to be drawn here is that naming decisions may be affected by a host of complex considerations that may not correlate in any simple way with ease of individuation.

Similarly, the decision to name an object by a count noun is generally governed by a number of complex criteria that may include the object’s function (e.g., a computer); the intentions of its creator (e.g., a painting, a collection); and its internal properties and characteristics, its essence [e.g., an animal; see Soja et al. (1992), Bloom (2001)]. As we have already pointed out, what licenses (and in fact guarantees) sameness of name is sameness of kind: we give the same verbal label to objects that fall in the same category. That is, our naming decisions typically depend on our decisions about kind membership.

What about the “shape bias?” Shape has been shown to be a central factor in the naming practices of both children and adults – and to be preferred over color, size, or texture as a basis of labeling objects with count nouns [Landau, Jones and Smith (1992), Landau, Smith and Jones (1992,1998), Landau and Leyton (1999)]. The question is whether attention to shape comes about as a consequence of learning count nouns, as the weak discontinuity proposals assume. Recall that, on this view, speakers of languages that possess count nouns come to map perceptual properties of objects (such as shape) onto count syntax as a result of their linguistic experience, and hence, develop a preference for extending count nouns to objects of similar perceptual contour.
Our position provides an alternative explanation for the fact that names for objects are often projected according to shape [see also Bloom (1994a, 2001) for related discussion]. Notice that, for several different reasons, humans are able to draw quick and correct inferences about object kind membership on the basis of perceptual contour. For artifact categories, shape is linked to function. Whether something is a shoe (and has the function of being worn on one’s feet) places constraints on its shape. In the case of natural kinds, such as plants, or animals, their shape is linked to their biological properties and evolutionary history: the giraffe’s neck is an obvious example. The usefulness and importance of shape for nonlinguistic cognition has been repeatedly emphasized in studies of nonlinguistic categorization [Rosch et al. (1976)] and studies of the visual representation of objects [Landau and Jackendoff (1993)].

If shape is a privileged cue to category membership for objects, it is no surprise that objects that have similar shapes (and probably belong to the same kind) also tend to share the same label. Expressions such as giraffe and shoe are typically extended to entities that share the same shape, because people use shape to categorize objects as giraffes or shoes, independently of language. If this is true, then attention to shape is not a product of learning count nouns (governed by a “shape bias”), but is a general cognitive preference that enters into categorization decisions.

Could it be that a shape bias, even though not quite accurate as a description of how adults label objects or individuals, determines naming decisions in children? Even 2- and 3-year-olds have been shown to be sensitive to this bias [Landau et al. (1992)]. It might be that children start out with this special procedure for extending count nouns to novel objects, and only later come to use more fine-grained cues to form (and label) object categories.

An obvious problem with this view is that children learn lots of count expressions for kinds that do not share shape. Preschoolers talk about grandmothers and telephones, animals and games. More generally, there is evidence that shape is not the only (or even the dominant) factor in early object labeling. When it contrasts with other important properties, it can be overridden. For instance, 4-year-olds know that a sponge shaped to look like a rock is in reality a sponge – and do not hesitate to call it so [Flavell, Flavell and Green (1983)]. Slightly older children recognize that membership to a kind (and labeling) is not determined by appearance only. A porcupine that has been transformed to look like a cactus is still a porcupine, and should still appropriately be called so [Keil (1989)]. Other studies have demonstrated the importance of intention in children’s naming decisions. When asked to draw a balloon and a lollipop, 3- and 4-year-olds typically produce similarly shaped drawings; nevertheless, they name these representations not on the basis of shape but according to what they intended them to depict [Bloom (2001)]. In another study in this series, 4-year-olds recognized that what a picture depicts does not depend only on what the picture looks like, but crucially depends on the creator’s intent (e.g., which object the person who drew the picture was looking at while drawing). As before, children’s naming decisions in these experiments go beyond perceptual features and incorporate deeper properties which determine kind membership.
If children name objects on the basis of kind membership, one would expect to see changes in the relative importance of shape during development, as children acquire increasingly sophisticated ways of generalizing object categories – and this seems to be true [Becker and Ward (1991), Imai, Gentner and chida (1994), Macario (1991)]. We conclude that it is not experience with language (more specifically, the acquisition of count nouns) that creates a shape bias for young learners; rather, attention to shape emerges as an accurate means of categorizing (and labeling) object kinds.

5. Conclusion

In the previous pages, we have surveyed several versions of the thesis that language may cause conceptual discontinuities – both across ages and across different speech communities. Specifically, we focused on the grammatical distinction between count and mass expressions, and we looked at how this distinction relates to our ontological concepts of objects and substances. We began by considering evidence against strong discontinuity proposals, according to which our ontological concepts are necessarily mediated by language. We next reviewed certain experimental results that have been taken as evidence for weakly discontinuous development of our folk ontology. While the results are certainly intriguing, neither the data as they stand nor the reasoning behind these studies have yet conclusively demonstrated language-related influences on the conceptualization of objects or material. Finally, we have argued that decisions to use material or shape as the basis for common labeling do not result from language-internal biases but are mediated by considerations of how material or shape impact decisions about kind membership.

Importantly, the data we surveyed revealed that humans – already from infancy – use individuation criteria that are more subtle than the object/substance distinction. Such criteria form a prime candidate for “core knowledge” (to use Spelke’s term), i.e., universal conceptual primitive notions that precede and structure language development. Both the precise form of these primitive notions and their relationship to the count-mass syntax offer fertile ground for future experimentation and theorizing. Results from such investigations will be important in further tracing the complex relationships between early linguistic and conceptual categories.

References

Ch. 11: Relations between Language and Thought


Li, P., and Y. Dunham (2005), “About things and stuff”, Poster to be presented at the Society for Research in Child Development Biennial Meeting, Atlanta, GA.


This page intentionally left blank
# Chapter 12

## DEFINITIONS IN CATEGORIZATION AND SIMILARITY JUDGMENTS

SERGE LAROCHELLE, DENIS COUSINEAU AND ANNIE ARCHAMBAULT

*Université de Montréal*

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>278</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>279</td>
</tr>
<tr>
<td>2. Importance rating and property selection</td>
<td>284</td>
</tr>
<tr>
<td>2.1. Method</td>
<td>284</td>
</tr>
<tr>
<td>2.1.1. Participants</td>
<td>284</td>
</tr>
<tr>
<td>2.1.2. Stimuli</td>
<td>284</td>
</tr>
<tr>
<td>2.1.3. Procedure</td>
<td>285</td>
</tr>
<tr>
<td>2.2. Results</td>
<td>286</td>
</tr>
<tr>
<td>3. Categorization judgments</td>
<td>288</td>
</tr>
<tr>
<td>3.1. Method</td>
<td>288</td>
</tr>
<tr>
<td>3.1.1. Stimuli</td>
<td>288</td>
</tr>
<tr>
<td>3.1.2. Procedure</td>
<td>290</td>
</tr>
<tr>
<td>3.2. Results</td>
<td>290</td>
</tr>
<tr>
<td>3.2.1. Experiment 1</td>
<td>290</td>
</tr>
<tr>
<td>3.2.2. Experiment 2</td>
<td>293</td>
</tr>
<tr>
<td>3.3. Discussion</td>
<td>296</td>
</tr>
<tr>
<td>4. Similarity judgments</td>
<td>297</td>
</tr>
<tr>
<td>4.1. Method</td>
<td>297</td>
</tr>
<tr>
<td>4.2. Results</td>
<td>298</td>
</tr>
<tr>
<td>5. General discussion</td>
<td>300</td>
</tr>
<tr>
<td>References</td>
<td>302</td>
</tr>
</tbody>
</table>
Abstract

Three experiments are reported that examined the role of definitions in categorization and similarity judgments. In each trial of the experiments, a verbal description was presented to the participants, along with a category name. Four different types of descriptions were used: Some asserted a set of properties deemed to define the category named, while others asserted characteristic properties. Some negated a set of properties deemed necessary for category membership, while others negated properties that were only characteristic of the category named. The results showed that necessary and characteristic properties are processed alike in both categorization and similarity judgments. Definitions and characteristic descriptions are also treated alike when making similarity judgments. However, definitions appear to be treated in a more unitary fashion when making categorization judgments. These results, which do not fit any of the main views of concepts (classical, probabilistic, or dual), are explained in terms of the “theory” theory.
1. Introduction

The research presented here bears on a venerable issue in cognitive science, namely whether concepts have defining properties or not. One methodology that has been used to experimentally address this question consists in asking participants to determine whether some entity does or does not belong to a specified category on the basis of a description of the entity. In some cases, the description contains information that is deemed to define the category, whereas in other cases, the descriptions lack such information. This study is based on a similar approach. One major novelty is that the amount of information in the descriptions also varies. The combination of these two factors, i.e., description length and description type, provides a fairly strong rationale for determining whether concepts are well defined or not. This rationale, like much research on categorization, is rooted in the classical view of concepts.

According to the classical view of concepts [Smith and Medin (1981), Komatsu (1992), Murphy (2002)], the mental representation of a category consists of properties that are common and unique to the objects in the category. As an example, the conceptual representation of the category labeled square would involve properties such as having four sides of equal length and four right angles. Since all squares are assumed to possess these properties, a figure lacking any one of these properties cannot be a member of the category square. Each of the above properties is therefore said to be individually necessary for category membership. Since squares are the only figures to have all of the above properties, any figure with these properties must belong to the category square. The above properties are therefore jointly sufficient for category membership. Sets of individually necessary and jointly sufficient properties are considered defining because they allow one to determine category membership in an all-or-none fashion. The rationale of the present research is based on the obvious, though unexploited, implication that the number of properties in such defining sets should not affect categorization behavior. Whether a definition involves one, five, or ten properties does not matter since objects need to have all the defining properties to belong to the designated category. Similarly, whether an object lacks one, five, or ten necessary properties does not matter either since the absence of a single necessary property is sufficient to reject category membership. Categorization behavior should also remain insensitive to the amount of characteristic information provided about an entity, so long as this information is insufficient to determine category membership.

These predictions are graphically illustrated in Figure 1. The abscissa represents the number of properties making up various types of descriptions. The ordinate shows the level of confidence that the described entity does or does not belong to a specified category, depending on the nature and size of the description involved. The line labeled $+X_d$ represents the performance that one would expect under the classical view when a set of $X$ defining properties is asserted ($+$) about the entity to be categorized. In this case, one should be absolutely confident that the entity described belongs to the specified category, irrespective of the number of properties in the definition. The line labeled $-X_n$ represents the performance that one would expect when $X$ necessary properties
are negated (−) with regard to the entity to be categorized. Figure 1 shows no effect of description size since the lack of only one necessary property should be sufficient to reject category membership with absolute certainty. Finally, the lines labeled +Xc and −Xc represent descriptions that either assert or negate X characteristic properties. Since these properties are neither necessary nor sufficient for category membership, they do not allow any absolute decision about the nature of the entity described (the two lines in Figure 1 have been separated only for illustrative purposes)1.

Very different predictions are obtained from a probabilistic view of conceptual representation. According to this view, the mental representation of categories involves properties that are common to some but not necessarily all members of a category. Such characteristic properties are also assumed to be involved in the categorization process. Although some probabilistic models have denied the existence of necessary and defining properties, such a strong stance is not essential [see Hampton (1998)]. What is crucial is that all types of properties, whether necessary, defining, or merely characteristic, be processed in the same way. While denying a special status for defining properties, the probabilistic view admits that some properties are more diagnostic than others and are

1 We do not know any contemporary researcher who would argue that non-necessary or nonsufficient characteristic information will be ignored, especially when it is the only information available for categorization. However, this prediction is useful for distinguishing the classical view from dual views of categorization, described later.
therefore weighed more heavily in categorical decisions. This is the case in models of the type described in Smith and Medin (1981) or proposed by Hampton (1979, 1998), for instance. Under a probabilistic view, categorization should therefore depend on the number and the diagnosticity of the properties asserted or negated with respect to an object. Predictions derived from a probabilistic view are shown in Figure 2. As shown, confidence in categorical decisions is now graded. Confidence increases with the number of properties that are asserted about the entity to be categorized and decreases with the number of properties that are negated. The increase is more pronounced for +Xd than for +Xc descriptions because properties that are common and unique to members of a category are more diagnostic of category membership and presumably weigh more heavily in categorical decisions than properties that are merely characteristic. Similar reasoning applies to the performance expected with −Xn descriptions, compared to −Xc descriptions. Since the former, but not the latter, are made up of properties that are thought to be common to all members of a category, they may weigh more heavily in categorical judgments, thereby causing a steeper change in confidence level with the number of properties negated.

Figure 3 shows the performance that one could expect under a dual view of categorization. The predictions illustrated for +Xd and −Xn descriptions are identical to those in Figure 1, being based on the assumptions that defining descriptions are processed holistically and that the negation of a single necessary property is sufficient to reject category membership. The predictions illustrated for +Xc and −Xc descriptions

![Probabilistic view](image)

**Fig. 2.** Predictions of a probabilistic view for descriptions made of X defining (d), necessary (n), or characteristic (c) properties that are either asserted (+) or negated (−).
are identical to those in Figure 2, being based on the assumption that characteristic properties individually contribute a certain weight to categorical decisions. Dual views of concept representation come in different guises. According to one proposal, characteristic properties are used mostly for the quick identification of objects, whereas true categorization must rest on a conceptual core of defining properties [Osherson and Smith (1981)]. To use a well-known example, the sight of a cookie-baking, gray-haired woman may lead to the hypothesis that the person is a grandmother. However, for true categorization, one must test whether the person is the mother of a parent. The instructions given to the participants in the experiments reported here emphasized true categorization.

A critical aspect of dual views is that defining properties must be distinct from non-defining properties, so as to be processed differently. This was the case in Smith, Shoben and Rips’ (1974) early model of category verification times. Defining and non-defining properties are also processed differently in McNamara and Sternberg’s (1983) mixed model. In the model, one mechanism checks for the presence of necessary or defining properties and a second one accumulates the evidence provided by characteristic properties. The latter type of property is assumed to determine categorization decisions when necessary or defining properties are absent, in which case the model’s decision sums the weight of the evidence provided by the characteristic properties. Therefore McNamara and Sternberg’s mixed model would make predictions very similar to those illustrated in Figure 3. Descriptions of the four types discussed so far (i.e., $+X_d$, $+X_c$, $-X_n$, and $-X_c$) are involved in Experiment 1.

![Dual view](image)

Fig. 3. Predictions of a dual view for descriptions made of $X$ defining (d), necessary (n), or characteristic (c) properties that are either asserted ($+$) or negated ($-$).
Testing dual views is complicated by the fact that some theories do not require all concepts to have both defining and characteristic properties. This is the case with psychological essentialism [Medin and Ortony (1989)], for instance. According to the theory, there is a deeply ingrained belief that objects have permanent, essential properties, which make them what they are, alongside their more variable and accidental properties. Such a belief in essence is assumed to apply to all natural kinds. For some categories, the belief in essence may be accompanied by a knowledge of the properties that are essential, or so intrinsically tied to the essential properties as to form a definition in the classical sense while, for other categories, the same person may hold more or less “inchoate” ideas [Medin and Ortony (1989), p. 184], or even rely on scientists to specify essential properties now or in the future [Putnam (1973, 1975)]. In the context of dual-view theories, an attempt to distinguish defining from nondefining properties must focus on categories for which subjects can provide defining properties, the question being whether they treat such properties differently from nondefining ones.

Another complication is that there are many different ways to integrate the information provided by descriptions. The predictions illustrated in Figures 1–3 are based on the assumptions that the characteristic properties have an additive effect on confidence level and that the expected change in confidence with every additional property is fairly constant. An alternative way to integrate the information would be to average the weight of the properties in the descriptions. Averaging rules are prevalent in information integration theory [Anderson (1981)]. In the field of categorization, such a rule has been proposed by Hampton (1988) to account for performance in experiments involving conjunctive concepts. In the present paradigm, performance would no longer depend on the number of properties in the description if the property weights were averaged instead of being summed. However, performance would still depend on the type of description. In short, the performance expected of probabilistic or mixed models based on an averaging rule could resemble that illustrated in Figure 1, thereby making the different views quite difficult to distinguish.

Such a distinction can be achieved, however, by incorporating conflicting information in the descriptions. Imagine a description asserting that an entity has all the defining properties of a specified category, but that it lacks a property that is characteristic of some members of the category. Such descriptions can be labeled +Xd−1c. According to the classical view, performance should not be affected by the absence of a characteristic property. The same would also be true of dual models, based on the assumption that characteristic information is ignored when a definition is present. However, even if one were to allow conflicting characteristic information to be taken into account, the effect on performance should still be independent of the number of properties in the definition, since definitions are assumed to have a holistic effect in classical and dual views. So, although the confidence level obtained with +Xd−1c descriptions might be smaller than that obtained with +Xd descriptions, it should not vary with definition size. By contrast, in a probabilistic model based on an averaging rule, the negation of a characteristic property should have differential effects on performance depending on the number of properties in the definition. The effect of the negated property on the mean weight should decrease as
the number of asserted properties increases and the denominator gets larger, thereby allowing confidence to increase with the number of defining properties in +Xd−1c descriptions. The same prediction applies to other types of descriptions containing conflicting evidence, such as +Xc−1c, −Xn+1c, and −Xc+1c. Not only does a probabilistic model based on an averaging rule still predict a change in confidence with description size, but the change in confidence level should be more pronounced for +Xd−1c than for +Xc−1c, and for −Xn+1c than for −Xc+1c, given the larger weight presumably associated with necessary and defining properties. In short, with descriptions containing conflicting information, models based on a sum vs. an average of weights generate a similar pattern of predictions, with the main difference between models residing in the linear vs. nonlinear shape of the description size functions. Descriptions involving conflicting information were used in Experiments 2 and 3, reported here.

2. Importance rating and property selection

One precondition for testing the hypotheses outlined in Section 1 is that participants must be able to identify properties deemed necessary for and defining of various categories. Another is that such properties must be considered to have more weight than merely characteristic properties. These two prerequisites were tested in the first two sessions of each experiment. The aim of the first session was to obtain the participants’ judgments concerning the importance of every property with respect to the associated category name. The goal of the second session, which had two parts, was to identify the properties that participants considered necessary (part 1), and necessary and sufficient (part 2). We will describe these two sessions, which were identical over all experiments, before turning to the results of the third session, which differed across experiments.

2.1. Method

2.1.1. Participants

The participants in this study were undergraduate students at the Université de Montréal, enrolled in departments other than Linguistics, Philosophy, or Psychology. Fourteen women and eight men, aged between 19 and 31, participated in Experiment 1. Thirteen women and three men, aged between 19 and 23, participated in Experiment 2. Sixteen other students (11 women and 5 men) participated in Experiment 3; their age ranged from 19 to 22. All were native speakers of French.

2.1.2. Stimuli

All the stimulus material was in French. The category names, which are listed in Table 1, were taken from an earlier study by Saumier (1993). Of the 24 category names listed, eight referred to artifacts, eight to biological kinds, and eight were assigned to a group
called “well-defined”\textsuperscript{2}. A list of properties was associated with each category name. These property lists were also taken from Saumier (1993), who obtained them by asking a group of undergraduate students to write down properties that describe the members of the specified category. No constraints were placed on the necessity, sufficiency, or even importance of the properties that the participants could give. The properties gathered by Saumier were generally used verbatim. The properties that were modified involved an indefinite quantifier over the members of a category. Since it would be the participants’ task to determine whether a property was true of all members of a category or not, we deleted all indefinite quantifiers from these properties, whether they were stated explicitly (e.g., “some are found in the countryside” for swallow) or implicitly through modals (e.g., “may have armrests” for chair).

2.1.3. Procedure

Participants took part in the first two sessions on separate days at least 1 week apart. Each session lasted between 1 and 2 h. These sessions were run using spreadsheet software with a separate file for each category. The category name appeared on top of the computer screen with each of its properties listed below on separate lines. The properties were listed either in alphabetical order or in reverse alphabetical order, depending on the participants and categories. Both orders were represented equally often over categories and participants.

Participants were told to interpret the category names literally rather than metaphorically, and to choose and stick to one meaning when they considered the category name to have more than one. For the first session, participants were instructed to rate the importance of each property with respect to the category named, using a scale ranging from 1 (extremely unimportant) to 9 (extremely important). Participants typed their

\begin{table}
\centering
\begin{tabular}{|l|l|l|l|}
\hline
\textbf{WELL-DEFINED TERMS} & \textbf{BIOLOGICAL KIND} & \textbf{ARTIFACT TERMS} \\
\hline
Célibataire ('Bachelor') & Carotte ('Carrot') & Avion ('Airplane') \\
Coup de circuit ('Home run') & Chien ('Dog') & Chaise ('Chair') \\
Deux ('Two') & Érable ('Maple tree') & Maison ('House') \\
Eau ('Water') & Hirondelle ('Swallow') & Marteau ('Hammer') \\
Grand-mère ('Grandmother') & Mouche ('Fly') & Ordinateur ('Computer') \\
Lundi ('Monday') & Pomme ('Apple') & Pantalon ('Pants') \\
Majorité ('Majority') & Truite ('Trout') & Revolver ('Revolver') \\
Triangle ('Triangle') & Tulipe ('Tulip') & Violon ('Violin') \\
\hline
\end{tabular}
\caption{Category names used in the experiments (with English translation)}
\end{table}

\textsuperscript{2} We chose the label “well-defined” because categories in the latter group were most likely to have a definition. However, there has been considerable debate about the existence, knowledge, and use of definitions for all the category types involved in our study.
ratings next to the corresponding properties. The order of presentation of the various category names and associated property lists was mixed over category types and differed across participants.

Subjects saw the same category names and associated property lists in the same order in session 2 as they had in session 1. However, the importance ratings gathered in session 1 were hidden from the participants. The first part of session 2 was devoted to the identification of properties deemed necessary for category membership. Participants were asked to distinguish properties considered to be “common to all category members” (by giving a rating of 2 to such properties), from those thought to be “present in some but not all members” (which were given a rating of 1). Finally, a rating of 0 was to be given to properties considered to be “present in none of the category members”\(^3\). Emphasis was placed on the fact that the properties listed could belong to members of other categories, but that this was irrelevant since the judgments required concerned only the category named.

The category names were once again presented in the same order in the second part of session 2 but, this time, the associated property lists were reduced to only those properties that had previously been identified as necessary by the participant. All prior ratings were also hidden from sight. Participants were asked to find the smallest set of properties that was also deemed “unique” to members of the named category. Emphasis was placed on the fact that each property in this set could be shared by members of other categories (provided that the set contained more than one property), but that the set as a whole had to be “true only” of the category named. Participants were asked to attribute a rating of 1 to all properties in the defining set selected and 0 to the remaining properties, attributing a 0 to all the properties listed if they did not find any defining set.

2.2. Results

As shown in Table 2, almost all participants identified at least one necessary property for almost all items. Ninety-five percent of the participants also succeeded in specifying a definition for the well-defined terms. The corresponding percentages were 86% for biological kinds and 92% for artifacts. Table 2 also gives the mean number of properties identified as necessary and defining for each type of category name involved in the experiments\(^4\). As shown, fewer properties were selected as necessary for well-defined terms than for artifact and biological kind terms. Table 2 shows that the definitions produced by the participants for the well-defined terms also contained a smaller number of properties than those concerning the other category types\(^5\).

\(^3\) The purpose of the 0 rating was to exclude properties deemed ludicrous or tied to a different meaning of the category name presented (e.g., triangle as a geometric figure or as a love situation).

\(^4\) These means were computed over participant–category pairs for which at least one necessary or defining property was identified. Since we were interested here in the size (and importance) of the sets of properties selected as necessary, and especially as defining, the data for participants who did not select a single necessary or defining property were excluded from the results.

\(^5\) Differences mentioned in the text are supported by statistical tests that were significant at least at the 0.05 level.
These results show that participants are quite confident not only of the existence of necessary and defining properties for various category names, but also in their knowledge of such properties, which was a prerequisite for testing the main hypotheses of the present study. Similar, though not identical, results had previously been obtained by McNamara and Sternberg (1983). Their subjects, like ours, identified necessary properties for almost all the test items in their study. However, their participants identified defining properties for only about 58% of the natural kind terms and 45% of the artifact terms, compared to close to 90% for both category types in our experiments. This difference is probably attributable to differences in procedure across the two studies. Participants in McNamara and Sternberg’s (1983) study did not have to select defining properties. Instead, they first had to select necessary properties from among those present on a master list, which also contained characteristic properties. Then, they had to select a set of sufficient properties from the same master list. The properties deemed defining for a given item resulted from the overlap between these two successive choices. By contrast, our participants were explicitly asked to identify a set of defining properties by selecting a set of sufficient properties from among those they had previously chosen as necessary. It may therefore have been easier for them to do so. Alternatively, one could argue that our procedure induced participants to identify sets of properties that were not always really thought to be sufficient. If this were the case, then categorization performance would be biased toward probabilistic predictions, which is contrary to the findings described below.

Table 2
Percentage (%) of participants who identified necessary or defining properties. Average number (N) and importance (Imp.) of the properties deemed necessary or defining for the well-defined, biological kind, and artifact terms over all experiments in the study

<table>
<thead>
<tr>
<th>Category type</th>
<th>Necessary</th>
<th></th>
<th>Defining</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>N</td>
<td>Imp.</td>
<td>%</td>
</tr>
<tr>
<td><strong>Well-defined</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp. 1</td>
<td>98</td>
<td>8.5</td>
<td>7.07</td>
<td>92</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>97</td>
<td>8.4</td>
<td>7.44</td>
<td>95</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>98</td>
<td>7.2</td>
<td>7.49</td>
<td>97</td>
</tr>
<tr>
<td>ALL</td>
<td>98</td>
<td>8.0</td>
<td>7.33</td>
<td>95</td>
</tr>
<tr>
<td><strong>Biological kind</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp. 1</td>
<td>100</td>
<td>13.0</td>
<td>7.02</td>
<td>81</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>99</td>
<td>13.0</td>
<td>7.02</td>
<td>93</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>98</td>
<td>10.2</td>
<td>7.42</td>
<td>85</td>
</tr>
<tr>
<td>ALL</td>
<td>99</td>
<td>12.0</td>
<td>7.15</td>
<td>86</td>
</tr>
<tr>
<td><strong>Artifact</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exp. 1</td>
<td>99</td>
<td>11.2</td>
<td>7.03</td>
<td>89</td>
</tr>
<tr>
<td>Exp. 2</td>
<td>100</td>
<td>11.4</td>
<td>7.14</td>
<td>95</td>
</tr>
<tr>
<td>Exp. 3</td>
<td>100</td>
<td>9.4</td>
<td>7.55</td>
<td>93</td>
</tr>
<tr>
<td>ALL</td>
<td>100</td>
<td>10.6</td>
<td>7.24</td>
<td>92</td>
</tr>
</tbody>
</table>
A second condition for testing the hypotheses stated in Section 1 is that necessary and defining properties must be given more weight than characteristic ones. Although not presented in Table 2, the importance ratings of the characteristic properties were lower than those of the necessary properties for all category names in all experiments. On average over all three experiments, characteristic properties received an importance rating of 5.43 for well-defined terms, 5.79 for biological kind terms, and 5.91 for artifact terms. In comparison, the average importance ratings for necessary properties were 7.33 for well-defined terms, 7.15 for biological kind terms, and 7.24 for artifacts. As shown in Table 2, ratings were even higher for the necessary properties that were selected to form a defining set. In short, the second condition for testing the effects of description type on categorization behavior, namely that necessary and defining properties be weighed more heavily than characteristic properties, appears to be met.

3. Categorization judgments

3.1. Method

The third and critical session of Experiments 1 and 2 was devoted to a categorization task. Participants were told that they would have to identify entities on the sole basis of the descriptions provided. A rating of 9 was used to indicate “absolute certainty” that the entity described was a member of the category named. At the opposite end, a rating of 1 would indicate “absolute certainty” that the described entity was not a member of the category named. Subjects were told that they did not have to identify the described entity at all costs. Emphasis was placed instead on the importance of giving the answer that was most logical or most likely to be true. A rating of 5 was to be given if the description did not allow the participant to decide whether the entity could or could not bear the name presented. Ratings between 5 and 9, and between 5 and 1, respectively, indicated an increasing certainty that the entity described belonged or did not belong to the category named.

3.1.1. Stimuli

The descriptions used in Experiments 1 and 2 were individually tailored to each participant and category name. Experiment 1 involved four different types of descriptions. +Xd descriptions were composed of the properties that had previously been identified by the participant (part 2 of session 2) as forming a classical definition for the corresponding category name. −Xn descriptions were composed of the negative version of properties that had previously been judged by the participant (part 1 of session 2) as being true of all members of the category. The necessary properties in −Xn descriptions

---

6 The negative versions of the properties were usually made by simply negating the main verb of the clauses. However, in order to prevent later decisions from being based solely on the surface form of the properties, antonyms were used whenever feasible (e.g., “is big” became “is small”; “is a man” became “is a woman”).
were chosen randomly, with the number of properties in such descriptions being the same as in the corresponding +Xd descriptions. In +Xc descriptions, a set of characteristic properties were asserted. These properties were randomly selected among those the participant thought to be true of only some members of the category. Finally, −Xc descriptions were formed of the negative version of randomly selected characteristic properties. The number of properties in the +Xc and −Xc descriptions was the same as in the corresponding +Xd and −Xn descriptions for a given participant and category name.

As mentioned in Section 1, distinguishing among views of categorization can be facilitated if one introduces conflicting information into the descriptions. This was done in Experiment 2, which involved eight different types of descriptions. Two types of descriptions contained the properties the participant considered to define a given category name, along with one or four properties thought to be only characteristic of the category. The defining properties were asserted and the characteristic properties were negated, whence the labels +Xd−1c and +Xd−4c. In two other types of descriptions, labeled +Xc−1c and +Xc−4c, different sets of characteristic properties were asserted and negated. The number of properties asserted was smaller than or equal to the number of properties the participant had deemed to be defining of the same category, and one or four other characteristic properties were negated. There were two description types, labeled −Xn+1c and −Xn+4c, in which a number of necessary properties, equal to the number of properties in the corresponding definition, were negated and one or four characteristic properties were asserted. Finally, in the last two descriptions, labeled −Xc+1c and −Xc+4c, a number of characteristic properties, smaller than or equal to the number of properties in the definition, were negated and one or four other characteristic properties were asserted.

Apart from the theoretical motivation sketched in Section 1, the use of descriptions containing conflicting information in Experiment 2 served a useful methodological purpose: they reduce the likelihood that flat description size functions in Experiment 1 result from ceiling (in the case of +Xd descriptions) or floor effects (in the case of −Xn descriptions). By moving the ratings away from the boundaries of the rating scale, the presence of conflicting characteristic information could make room for the ratings to increase with the number of defining properties and decrease with the number of necessary properties. Failure to find such changes in ratings with description size in Experiment 2 would constitute further evidence for the classical or dual view, and so would the finding that confidence levels are not affected by the number of conflicting characteristic properties.

In both experiments, descriptions were created only for category names that had been assigned a definition by the participant. Even for such items, it was not always possible to construct all the required types of descriptions. In some cases, there were too few necessary properties left to negate. In others, there were too few or no characteristic properties to assert or to negate. Participants were tested with items for which all the matched description types could be made. In Experiment 1, each participant was also tested with the descriptions constructed for another, randomly selected subject. These other descriptions were intended as fillers. Participants in Experiment 2 saw only descriptions derived from their own property selection protocol.
3.1.2. Procedure

The third session took place a week to a month after session 2 and lasted between 90 and 120 min. The test order of the various category names and description types was randomized in both experiments. In each trial of Experiment 1, a description first appeared on top of the screen, one property per line in random order. Participants were instructed to read the description and then press the space bar. This caused a category name to appear below the description, along with the rating scale. Participants were warned that their task was not to guess the category name that would appear. Rather, their task was to give the most logical response considering the description and category name presented. Participants indicated the chosen rating by typing the corresponding digit on the keyboard. In Experiment 2, the descriptions and category names were displayed simultaneously on the screen, along with the rating scale. Asserted and negated properties were randomly mixed on the screen.

3.2. Results

3.2.1. Experiment 1

A total of 1768 descriptions were generated for the 22 participants in Experiment 1. Since construction of these descriptions was not fully automated, errors were unavoidable. All descriptions found to contain an error were excluded from the analyses. Also excluded were the descriptions that matched a description containing an error. For instance, if an error was found in the +Xd description of a given participant for a given category name, the matching +Xc description was also excluded from the data, and vice versa. The same was done for −Xn and −Xc descriptions. This caused the loss of about 2.0% of the data, with a total of 1736 descriptions remaining for all participants, category names, and description types. In short, the data used to contrast performance obtained with descriptions containing only characteristic properties versus those containing defining or necessary properties came from the same participant–category pairs.

Since the size of +Xd descriptions (i.e., definition size) had been determined by the participants in part 2 of session 2 and, since it determined the size of the matching +Xc, −Xn, and −Xc descriptions, there was little chance that all conditions of the experiment would be equally represented under all description sizes. Over all participants and category names, description size varied from 1 to over 10 propositions⁷. Not all category

⁷ The properties in the descriptions were sometimes composed of more than one proposition. If propositions are defined as the smallest units of meaning the truth of which can be verified, then one must consider the property “has a long stem” (tulip), for instance, to be composed of two propositions: “has a stem” and “stem is long.” Measuring description size in terms of number of propositions instead of number of properties therefore provides a more precise measure of the amount of information in the descriptions. The results presented here are based on the number of propositions as the measure of description size. However, analyses of categorization performance based on number of properties yielded results very similar to those presented.
types were represented among descriptions of size 8 or more. Even for description sizes of less than 8, there was no guarantee that all category names would contribute observations, or an even number of observations. This uneven distribution of observations over description size led us to compute a separate linear regression for every category name and description type, using the number of propositions in the descriptions as the predictor variable and the categorization ratings provided by different participants as the predicted variable. The slopes of these regressions, as well as other measures, were then entered into ANOVAs, using category name as the random factor. Category type was considered as a “between” factor, and description type was considered as a “within” factor. Separate ANOVAs were computed on the results obtained with positive descriptions (+Xd vs. +Xc) and with negative descriptions (−Xn and −Xc). Note that our use of regression coefficients as a dependent variable in the ANOVAs does not imply that the description size effect is strictly linear, or even that it has a significant linear component. Slopes are simply used as a measure of the description size effects, one that takes into account all the available data without requiring that every description size be equally represented within each category name.

Figure 4 summarizes the results obtained in Experiment 1. A separate panel is devoted to each category type. The height of the four points in each panel corresponds to the mean rating, computed over category names, for each description type. The location of the points along the abscissa corresponds to the average number of propositions. The four lines in each panel were drawn using the average intercepts and slopes computed over category names for each description type. The main result obtained with positive descriptions is that slopes were generally shallower for defining descriptions (mean = 0.08) than for descriptions made of characteristic properties (mean = 0.21). This convergence of the regression lines for +Xd and +Xc descriptions did not differ significantly across category types. Turning to negative descriptions, Figure 4 shows some convergence of the regression lines for −Xn and −Xc descriptions. However, the statistical analyses did not reveal the slopes obtained with −Xn (mean = −0.12) and −Xc (mean = −0.19) to differ from each other for any category type.

The mean rating obtained with +Xd descriptions (8.29) was larger than that obtained with +Xc descriptions (6.19). The difference in mean rating between these description types did not vary significantly across category types. The same was true of the difference in mean rating between −Xn and −Xc descriptions. However, in the case of negative descriptions, the mean rating obtained with descriptions containing necessary properties (1.94) was smaller than that obtained with descriptions made up of characteristic properties (2.79). An interesting finding is that the mean ratings obtained with −Xc descriptions were about twice as far from the center of the scale (5) as those obtained with +Xc descriptions. This result suggests that negating characteristic properties of an entity has more influence on categorization than asserting similar properties or, perhaps more generally, that positive and negative evidence are weighed differently in categorical decisions.

Although the pattern of results obtained in Experiment 1 and shown in Figure 4 appears quite similar to that predicted by some dual models, especially in the case of
Fig. 4. Mean categorization ratings, along with the mean intercepts and slopes obtained with the various description and category types (Experiment 1).
well-defined terms, the conclusions that are statistically supported are the following: First, the description size effects obtained with $+Xc$ and $-Xc$ descriptions show that characteristic information is taken into account when no other information is available for categorization (as just about everyone would have expected). Second, the lack of difference in slopes between $-Xn$ and $-Xc$ descriptions shows that necessary properties may not be processed differently from characteristic properties, although the difference in mean ratings shows that necessary properties weigh more heavily in categorical decisions. The third finding, which is somewhat surprising in light of the previous one, is that defining properties not only appear to weigh more than characteristic properties, as evidenced by the larger average ratings obtained with $+Xd$ descriptions, but they also appear to be processed in a more unitary fashion, as indicated by the shallower slopes obtained with $+Xd$ descriptions. There was no reliable evidence that participants behaved differently depending on the type of category involved.

As mentioned earlier, the results of Experiment 1 could have been contaminated by ceiling and floor effects of a methodological nature. This would have been especially likely to occur for $+Xd$ and $-Xn$ descriptions since the mean ratings obtained with such descriptions were quite close to the extremities of the rating scale. Experiment 2 was less sensitive to this possible bias.

3.2.2. Experiment 2

As in Experiment 1, the data used to contrast performance obtained with descriptions containing only characteristic properties versus those also containing defining or necessary properties came from the same participant–category pairs. This was achieved by eliminating every description that contained an error, along with its matched description (e.g., $+Xd-1c$ and $+Xc-1c$). Such error elimination left a total of 2482 trials, for an average of over 150 per participant. To ease comparison with the results of Experiment 1, description size was measured by counting the propositions forming the variable parts of the descriptions (i.e., the $+Xd$, $-Xn$, and $+/−Xc$ parts) and ignoring the propositions forming the constant parts (i.e., the $+/−1c$ and $+/−4c$ parts). We will present the results obtained with one and four conflicting characteristic properties together, noting differences when they occur. We will call “positive” the descriptions that asserted a variable number of defining ($+Xd$) or characteristic properties ($+Xc$) and drop the $−1c$ and $−4c$ from the labels for the conditions, except when needed. Similarly, the descriptions that negated a variable number of necessary or characteristic properties will be called “negative” and labeled $−Xn$ and $−Xc$, despite the assertion of one or four characteristic properties.

Figure 5 summarizes the results in Experiment 2, averaged over both conflict conditions. The format of the figure is the same as that of Figure 4. The most obvious result is that performance was affected by the presence of conflicting characteristic information. The mean rating obtained with $+Xd−4c$ descriptions (5.11) was smaller than that obtained with $+Xd−1c$ descriptions (7.20), which in turn was smaller than that obtained with $+Xd$ descriptions in Experiment 1 (8.29). This result shows that characteristic
Fig. 5. Mean categorization ratings, along with the mean intercepts and slopes obtained with the various description and category types (Experiment 2).
properties are taken into account, even in the presence of defining properties. Descriptions made of characteristic properties were also affected by conflicting information. The mean rating obtained with $+Xc-4c$ descriptions (3.21) was smaller than that obtained with $+Xc-1c$ descriptions (5.01), which in turn was smaller than that obtained with $+Xc$ descriptions in Experiment 1 (6.19). The overall difference in mean rating between $+Xd$ descriptions (6.16) and $+Xc$ descriptions (4.11) was significant in Experiment 2, as it was in Experiment 1. Category type did not have any effect on mean ratings, nor did it interact with any other factor in the analysis.

Given that the mean ratings obtained with positive descriptions in Experiment 2 were smaller than those obtained in Experiment 1, the slopes should be less affected by ceiling effects. The average slope obtained with $+Xd$ descriptions in Experiment 2 (0.22) was indeed greater than that obtained in Experiment 1 (0.08). Nonetheless, the slopes for $+Xd$ descriptions in Experiment 2 remained shallower, on average, than those obtained with $+Xc$ descriptions (0.45). Description type did not interact with the number of conflicting properties, as the mean slopes obtained with one or four conflicting properties were very similar for $+Xd$ descriptions (0.23 and 0.22, respectively) and for $+Xc$ descriptions (0.48 and 0.42, respectively). Figure 5 shows the convergence of the regression lines for $+Xd$ and $+Xc$ descriptions to be less pronounced for artifacts than for the other category types, which resulted in a marginally significant Description X Category-type interaction.

Turning to negative descriptions, those containing necessary properties yielded a smaller average rating (2.79) than descriptions composed only of characteristic properties (4.00). Both types of descriptions were affected by the amount of conflicting information. The mean rating obtained with $-Xc+4c$ descriptions (5.15) was larger than that obtained with $-Xc+1c$ descriptions (2.85); the latter was similar to the rating obtained with $-Xc$ descriptions in Experiment 1 (2.79). Similarly, the mean rating obtained with $-Xn+4c$ (3.64) descriptions was larger than that obtained with $-Xn+1c$ descriptions (1.94), but the latter was equal to the rating obtained with $-Xn$ descriptions in Experiment 1. Given the small ratings obtained, floor effects were still possible with $-Xn+1c$ descriptions in Experiment 2. Category type had a close to significant effect on mean ratings, due to the somewhat larger ratings obtained with well-defined terms. However, category type failed to interact with any other factor in the analysis.

As in Experiment 1, the mean slope obtained with $-Xn$ descriptions ($-0.13$) did not differ significantly from that obtained with $-Xc$ descriptions ($-0.17$). However, the slopes for $-Xn$ descriptions were affected by the number of conflicting properties, being smaller with one conflicting property ($-0.05$) than with four ($-0.22$). As mentioned, the shallow slopes obtained with $-Xn+1c$ descriptions could still be due to a floor effect. This possibility is supported by the fact that the slopes for $-Xc+1c$ ($-0.16$) and $-Xc+4c$ ($-0.17$) were almost identical and comparable to those obtained with $-Xn+4c$ descriptions. In short, apart from some possible floor effects, there was no evidence of any convergence in the regression lines obtained with $-Xn$ and $-Xc$ descriptions for any of the category types.
Finally, it may be worth noting that negative descriptions generally produced more extreme ratings than positive descriptions, which suggests once again that positive and negative evidence is not weighed equally in categorical decisions.

3.3. Discussion

The results of Experiment 2 are generally in line with those of Experiment 1. However, Experiment 2 extends the findings of Experiment 1 by showing that characteristic information is taken into account in categorization judgments, even when necessary or defining information is present. This result is clearly at odds with a strict classical view. The results of both experiments also conflict with a purely probabilistic view. Probabilistic models incorporating a summing rule predict more pronounced description size effects for $+X_d$ and $-X_n$ descriptions than for $+X_c$ and $-X_c$ descriptions, respectively. Probabilistic models based on an averaging rule make similar predictions for description incorporating conflicting information, although the differences in slopes would not be pronounced as in a summing model. In fact, none of the category types produced larger description size effects for descriptions containing necessary or defining properties than for descriptions made only of characteristic properties.

The results fit a dual view of concept representation best, although the fit is not perfect. Although the slopes obtained with $+X_d$ descriptions in Experiment 2 were greater than zero, they were still much smaller than those obtained with $+X_c$ descriptions, which suggests that definitions were processed in a more holistic, though not perfectly unitary, fashion. Since the analyses reported here concerned only items for which the subjects could provide a definition, the differential size effects obtained with $+X_d$ and $+X_c$ properties suggest that participants did distinguish between defining and nondefining properties in the representation of the same items, which is the essence of a dual view. A dual view can also accommodate the effects obtained with characteristic properties inasmuch as it is assumed that such information is taken into account, in the presence as well as in the absence of necessary or defining information. The main difficulty for a dual view is that $-X_n$ and $-X_c$ properties did not seem to be processed differently, as both conditions exhibited similar description size effects. Smith and Sloman (1994) suggested that rule-based categorization might be restricted to situations where no characteristic information is available. Since such information was present in the descriptions used in Experiment 2, their proposal is consistent with the lack of convergence obtained with $-X_n$ and $-X_c$ descriptions but, by the same token, it fails to account for the convergence obtained with $+X_d$ and $+X_c$ descriptions. The results obtained with $-X_n$ description in Experiments 1 and 2 are also inconsistent with McNamara and Sternberg’s (1983) mixed model.

Murphy and Medin’s (1985) “theory” theory might account for the pattern of results obtained. Their approach rejects the notion that concepts are represented by lists of isolated features or properties. Instead, the properties of a concept are thought to be

---

8 The same is true of models involving a multiplicative rule, such as those found in exemplar models.
interconnected by a complex web of relations — foremost among which are causal relations — that provide an explanatory theory of what it takes to be a member of a category. As Pothos and Hahn (2000) have suggested, such an explanatory theory can serve as the basis for selecting properties that are considered necessary and/or sufficient for category membership. Suppose that this was the case in our experiments. By virtue of being based on an underlying theory, the properties selected as defining would have been more critical for category membership than those considered to be only characteristic. So defining properties would have had more weight than characteristic properties. Moreover, since the entire set of properties considered to be defining by the participants were included in +Xd descriptions, these descriptions would have preserved whatever relationships the participants found to exist among these properties. In other words, the descriptions made up of defining properties would have maintained the coherence of the underlying theory. By contrast, the descriptions made up of characteristic properties would have had little conceptual cohesion since the constituting properties were selected randomly. This difference in cohesiveness would explain why descriptions made up of defining properties were treated more holistically than descriptions made up of characteristic properties. With +Xc descriptions, the lack of conceptual cohesion may have left the participants little choice but to process each property separately. The properties selected as necessary by the participants were also presumably more critical for category membership than those selected as characteristic. However, the necessary properties entering into −Xn descriptions were selected randomly. More importantly, they were negated. As a result, −Xn descriptions could only violate the theories of what it takes to be a member of a category, which would explain why such descriptions were processed in the same piecemeal fashion as −Xc descriptions.

The main evidence quoted in favor of the “theory” theory comes from dissociations between categorization and similarity judgments [Rips (1989), Rips and Collins (1993)]; the idea is that the former, but not the latter, rely on a theory of what it takes to be a member of a category. A typical example, is that of a gray-haired woman, who may look very much like a grandmother, but who cannot be one if she is not the mother of a parent, because this is essential to the theory of grandmotherhood. If similarity judgments need not rely on a theory, then the descriptions containing defining properties should be processed in the same way as descriptions made up of characteristic properties. In other words, the description size effects obtained with +Xd descriptions should be similar to those obtained with +Xc descriptions. Testing this hypothesis was the goal of Experiment 3.

4. Similarity judgments

4.1. Method

In Experiment 3, subjects had to judge how similar the described entity was to members of the category named. The ratings were done on a 9-point scale, from “extremely
dissimilar” (rating of 1) to “extremely similar” (rating of 9). Using each individual participant’s property selection protocol, we constructed the same eight different types of descriptions as in Experiment 2. Except for the nature of the judgment required, the procedure was identical to that of Experiment 2. Error elimination left a total of 2532 trials.

4.2. Results

Figure 6 shows the mean ratings, averaged over both conflict conditions, along with the best-fitting average regression lines. The main new finding of Experiment 3 is that description type failed to have a significant effect on the slopes obtained with positive descriptions, the mean slopes being 0.31 and 0.37 for +Xd and +Xc descriptions, respectively. The interaction between description and category type failed to reach significance, and so did the main category-type effect. In short, similarity judgments seem to have been affected by the size of definitions about as much as they were by the size of descriptions containing only characteristic descriptions. The number of conflicting properties had no effect on the slopes, nor did it interact with category type or description type. To test the reliability of differences between categorization and similarity judgments, we compared the slopes obtained in Experiments 2 and 3. There was no overall difference across experiments in the slopes obtained with positive descriptions but judgment type did interact with description type, this interaction being the sole significant one in the ANOVA (all other values of $p > 0.10$).

The mean ratings obtained in Experiment 3 were affected by the amount of conflicting information in the descriptions. The mean similarity rating obtained with +Xd−1c descriptions (7.15) was larger than that obtained with +Xd−4c descriptions (4.88), and so was the mean rating obtained with +Xc−1c descriptions (5.38), compared to that obtained with +Xc−4c descriptions (3.31). The overall difference between +Xd descriptions (6.01) and +Xc descriptions (4.34) was significant. Compared to the categorization ratings obtained in Experiment 2, the similarity ratings obtained with +Xd descriptions were slightly smaller on average (6.02 vs. 6.16), while those obtained with +Xc descriptions were slightly larger (4.34 vs. 4.11), which resulted in a significant interaction between judgment and description type. There was no other interaction involving judgment type in the analysis comparing the ratings obtained with positive descriptions in Experiments 2 and 3.

The mean ratings obtained with negative descriptions were more extreme than those obtained with positive descriptions, as usual. However, they were also affected by the number of conflicting properties. The mean similarity rating obtained with −Xn+1c descriptions was 2.17, compared to 4.44 for −Xn+4c. The mean similarity ratings for −Xc+1c and −Xc+4c were 3.08 and 5.52, respectively. The presence of a three-way interaction involving category type indicates that there were variations across categories in the relative resistance of −Xn and −Xc descriptions to the amount of conflicting information. Despite this interaction, −Xn descriptions always produced smaller ratings (mean = 3.31) than −Xc descriptions (mean = 4.30). Comparing the similarity ratings of Experiment 3 with the categorization ratings of Experiment 2 revealed that the former
Fig. 6. Mean similarity ratings, along with the mean intercepts and slopes obtained with the various description and category types (Experiment 3).
were less extreme. The differences in mean ratings across experiments were more pronounced with descriptions containing four conflicting properties, than with descriptions containing only one such property. This was especially true of −Xn descriptions, so that performance obtained with −Xn+1c could still be affected by a floor effect.

As shown in Figure 6, the slopes obtained with −Xn descriptions did not differ, on average, from those obtained with −Xc descriptions. However, these mean slopes hide an interaction with the number of conflicting properties, similar to that found in Experiment 2. The slopes obtained with −Xn descriptions were affected by the number of conflicting properties; the mean slope for −Xn+1c descriptions was −0.12 versus −0.42 for −Xn+4c descriptions. There was little variation in this pattern of effect over categories. The shallow slopes obtained with −Xn+1c descriptions might still reflect some floor effect, given that the mean rating was still very low. By contrast, the slopes obtained with −Xc+1c (mean = −0.26) and −Xc+4c (mean = −0.24) descriptions were almost identical, but there were some variations in the size of the difference across category types. These variations caused a significant three-way interaction involving category and description type along with number of conflicting properties. Since there was no such interaction in the slopes obtained with negative descriptions in Experiment 2, the ANOVA comparing the results of both experiments showed a significant four-way interaction.

5. General discussion

Let us first briefly review the main findings of this study, in decreasing order of certainty. The first finding is that the weight of information on categorization and similarity judgments depends on whether the properties are asserted or negated with respect to the entity concerned. Negated properties, whether necessary or merely characteristic, had a greater influence on the average categorization and similarity ratings than properties that were asserted, be they defining or characteristic. The second finding is that some properties that are identified as common to all category members are still not treated as necessary for category membership. Hampton (1995) made valiant efforts to construct descriptions comprising properties that would be considered necessary for category membership by the participants in the study. His repeated failures, along with our findings, suggest that no property is considered as such. Rips’ (1989) study is often quoted as evidence against this conclusion, but Rips’ results were obtained in very limited conditions and, even in similar conditions, it has been difficult to replicate them [see Smith and Sloman (1994), Pothos and Hahn (2000)].

The third finding is that definitions seem to have a special status in categorization. This finding is perplexing from a classical perspective: How can definitions receive special treatment when the necessary properties that compose them do not? One possible explanation of this result is that people rely more on sufficiency than on necessity in categorization. The problem with such an interpretation is it requires two different notions of sufficiency: Why would the assertion of a set of unique properties be “sufficient” to accept category membership while the negation of common properties would
not be “sufficient” to reject category membership? This explanation is also problematic from a computational point of view, since testing necessity seems less costly than testing sufficiency. In terms of processing, it can certainly be arduous to determine whether a given property applies to all members of a category when the category is large; nevertheless, at least this is a finite task. By contrast, determining whether a set of properties is unique to the members of a category requires one to consider an indefinite number of alternative categories, which can be an endless task unless the decision can be reached deductively instead of inductively. For this reason, we prefer to attribute the results obtained with definitions to their conceptual coherence. If this interpretation is correct, then explanatory coherence does not seem to play the same role in similarity judgments. The fact that the similarity ratings were not as affected by the nature of the properties in the descriptions is a fourth important finding of our study.

We have argued that descriptions containing necessary properties lacked conceptual coherence since these properties were negated with respect to the entities to be categorized. Alternately, one could argue that these descriptions lacked cohesion because the necessary properties, like the characteristic properties, were selected randomly. An easy way to discriminate between these possibilities would be to negate all the necessary properties in the definitions. Such descriptions would have as much surface cohesion as the corresponding definitions, but they would lack the underlying conceptual coherence. Note that the outcome of such a test would not invalidate our previous conclusion about the lack of necessary properties. A set of truly necessary properties should still be treated as such, even when they are selected randomly.

The results reported so far concern the effects of description type and size on categorization and similarity judgments. They are based on a rather indirect measure, namely regression coefficients. Our interest in description size stemmed from the fact that different theories and models of categorization make different predictions about its effect on performance. However, the critical aspect of these theories and models does not reside in description size per se, but in the nature of the rules used to integrate the available information. Regression coefficients allowed us to test some of the models’ predictions, but they precluded our addressing others, namely those having to do with the rule(s) used to combine the information provided by the descriptions. Since many models differ in precisely this respect, we complemented the qualitative, hypothesis-testing approach described so far with a quantitative, model-fitting approach.

The models described in Section 1 were fit to the same categorization or similarity ratings involved in the previous analyses. However, instead of using a derivative measure of performance as the predicted variable, the models were fit to the raw categorization and similarity ratings obtained in the experiments. The predictor variable was derived from the importance ratings obtained in session 1 of the experiments for each the properties entering into the descriptions. The predictor variable was therefore adapted to each participant, category, and description used. In some probabilistic models, the importance ratings collected were simply added or subtracted depending on whether the corresponding property was asserted or negated in the description. In other models, the weights were then averaged over the number of properties in the descriptions. We
also attempted to fit the results, using multiplicative rules. Such single-process models were compared with dual-process models involving special treatment of the necessary and/or defining properties. This special treatment usually consisted in attributing maximum weight to descriptions containing such information, irrespective of the number of properties in the descriptions and of the presence or absence of conflicting information.

One of the best accounts of the categorization and similarity judgments was obtained by assuming that, in all but one circumstance, participants simply averaged the weights of the properties in the descriptions. The one exception concerned definitions, to which the model assigned maximum weight in the categorization task, but not in the similarity judgment task. This model accounted for close to half of the variance over all experiments. This is not negligible considering that the model was fit to over 6000 data points, involving both subject and item variability, using only two estimated parameters: the intercept of the regression of obtained over predicted (categorization or similarity) ratings was 4.17 and the slope, 0.36. Although this model did better than many others, some of the models tested did about as well. So, on the basis of available evidence, it appears premature to commit oneself to a specific integration rule, whether it be the summation rule proposed by McNamara and Sternberg (1983), an averaging rule such as that proposed by Hampton (1988), or a multiplicative scheme [Hampton (1995)].

Another issue that remains open is whether some types of categories are represented and processed differently from others. There was some indication in the categorization data that definitions may not have been treated as differently from characteristic descriptions when they concerned artifact names as when they concerned well-defined and biological kind names. However, the lack of any reliable category-by-description-type interaction prevents rejection of the null hypothesis. One should be equally wary of asserting that all categories were treated alike since the statistical analyses performed on the regression coefficients were not very powerful, involving only eight items per category type.

Finally, it must be stressed once again that the results presented here were obtained in conditions that gave the participants ample time to reach a decision. Such conditions make our interpretation of the results in terms of underlying theories quite plausible. By the same token, they impose caution about assuming that similar results would obtain in quick perceptual identification or online language processing.

References

This page intentionally left blank
Chapter 13

WHY (MOST) CONCEPTS AREN’T CATEGORIES

RUTH GARRETT MILLIKAN

University of Connecticut
Copyright Ruth Garrett Millikan 2005

Contents

Abstract 305
1. Introduction 306
2. Species are not categories 306
3. Three kinds of (Aristotelian) “substances” 307
   3.1. Historical kinds 307
   3.2. Eternal kinds 308
   3.3. Individuals 309
4. Concepts of individuals 310
5. Concepts of substances more generally 311
6. Substances encountered through language 312
References 315

Abstract

In the last century, hundreds of experiments were conducted by psychologists trying to discover how people “classify” or “categorize” items under kind (category) words such as dog, chair, and fruit. The position I have taken about such words is that they do not designate classes but rather units of an entirely different kind. A very few, uncompounded nouns that designate classes do exist, but words like dog, chair, and fruit are not among them. In this chapter I introduce this contrary viewpoint. I cannot attempt to defend it at any length in this short essay, but I will present a portion of this view, using the most intuitively understandable terms I can muster. The details are spelled out in On Clear and Confused Ideas (Millikan, 2000), which I will refer to as OCCI.
1. Introduction

One place to begin this discussion is with the claims by biologist Ghiselin (1974, 1981) and philosopher David Hull (1978) about what biological species really are. They maintain that to be members of the same species, individual animals must belong to historical lineages that have a common origin. They do not have to be similar to one another in any specified way. For example, there are no genes that every dog has in common with every other dog: every dog gene has alternative forms (alleles). Similarly, there are no properties that every dog has in common with every other dog. Nor is it mere overlap in properties or resemblance to some paradigm that makes a group of dogs conspecific. Highly similar species that have different historical origins do not form one species, but rather form several species. According to Ghiselin and Hull, species are not classes that share a similarity, but rather are big, scattered, historical individuals enduring through time. Species are entities somewhat like the Kennedy family, which is held together not by “family resemblance” in Wittgenstein’s sense but by blood ties.

On the other hand, in the case of species, blood ties are bound to be accompanied by considerable overlap in properties. The reason for this overlap is that if the species reproduces asexually, the progeny are clones. Overlap occurs if the species reproduces sexually, since each gene in the gene pool has to fit with a random selection of other genes from the pool, to help produce a viable individual frequently enough so that it can get itself reproduced often enough to not be eliminated from the gene pool. No single gene that changes the animal in very extreme ways can survive. This results in what is called “homeostasis” in the gene pool. Thus, the various individuals within a species mostly resemble one another in a great variety of ways, but do not all resemble one another in any particular ways. What pulls them together as a group is not just that they have common or overlapping properties, but that they tend to have common and overlapping properties for a good reason. One individual is like the next for a good reason. There is a good explanation of why one is likely to be like the next. Various kinds of inductions drawn about the members of a species are likely to be sound, owing to certain kinds of causal connections among these members.

In this analysis of what pulls the members of a species together, species are not classes. Classes are defined by the members having certain common properties. Fuzzy classes may be defined by the members having overlapping properties, or by their having many properties in common with a paradigm, or paradigms. But the members of a class do not need to be like one another, for any reason. They may be like one another quite by accident. Categories are classes or fuzzy classes. Species names are not names of categories.

2. Species are not categories

I will explain first, why this point – that species names are not names of categories – is important and second, how this can be generalized. It is important because it explains why it is possible to study a species as such, and to gather stable information about it.
If there is a reason why one dog is likely to be similar to the next dog in a good number of respects, then there is a reason why studying one dog is likely to yield a considerable amount of probable knowledge about the next dog. In fact, of course, it is possible to learn a great deal about dogs. Consider how much time might be spent on this by a student at a veterinary college. True, all this knowledge is merely probable knowledge. Whatever one learns about the properties of dogs, it will not be analytic or necessary that every individual dog has each of those properties. But mere classes are not things one can learn anything at all about by induction. If, given one member of a class, there is no reason why the next member is likely to be similar to it, then if any inductions about the class turn out true conclusions, this can only happen by accident. For example, it seems likely that there is no reason why one red triangular object should tend to be like the next one in any aspect other than redness and triangularity, therefore, it is not likely that discovering, say, that one red triangular object is sweet will be of any use in predicting the taste of the next one.

The way in which dogs are cemented into a unit is important, then, because it is only when individuals are cemented into a unit in some analogous way, such that there is a reason why one individual should be like the next, that we can obtain knowledge about this unit, unless, of course, by examining every member separately. It is obvious, then, why this sort of unit tends to acquire a name. Names for mere classes are, in most contexts, quite useless. Names for units of this kind are names for the seeds on which all empirical knowledge is built, for all empirical knowledge is inductive.

The point about dogs being cemented into a unit is generalized by noticing other kinds of relations that tend to cement a unit together, such that there is a reason why one individual within the unit, or one part of the unit, is liable to be like another. I have a name for units of this general kind, taken from Aristotle. Call them “substances,” a word which non-philosophers may need to read as a new technical term, but which philosophers may recognize as fairly traditional. Aristotle spoke of “secondary substances” – the unit dog would be an example – and of “primary substances,” which were individuals. (Recall that Ghiselin and Hull said that dogs were big, scattered, historical individuals enduring through time.) There is a chapter in OCCI detailing many kinds of substances. Here I will discuss only a few kinds of substances, but enough, perhaps, to give you a taste of this topic.

3. Three kinds of (Aristotelian) “substances”

Aristotelian “substances” fall roughly into at least three basic sorts, which I call “historical kinds,” “eternal kinds,” and “individuals.”

3.1. Historical kinds

A “historical kind” (like dogs, for example) is a collection of individuals scattered over a definite spatiotemporal area; the individuals are causally related to one another in
such a manner that each is likely to be similar to the next in a variety of aspects. There are two most obvious sorts of things that cause members of a historical kind to be like one another. First, something akin to reproduction or copying has been going on, all the various individuals having been produced from one another or from the same models. Second, various members have been produced by, in, or in response to, the very same ongoing historical environment – for example, in response to the presence of members of other ongoing historical kinds. Third – an ubiquitous causal factor often supporting the first factor – some “function” is served by members of the kind, where “function” is understood roughly in the biological sense as an effect raising the probability that its cause will be reproduced. It is typical for several of these kinds of causes to be combined. Artifacts are often good examples of this.

Consider, for example, chairs. Chairs have been designed to fit the physical dimensions and practical and aesthetic preferences of humans, who are much alike in relevant respects for the same reasons dogs are. Moreover, the design of a chair is pretty invariably influenced by the design of previous chairs, typically because these previous chairs have functioned well and were aesthetically pleasing within a cultural setting, and typically, relevant aspects of the cultural setting were reproduced elements as well. For these reasons, chairs form a rough “historical kind.” There are historical reasons, which have nothing to do with any arbitrary points of definition, why one knows roughly what to expect when someone offers to bring a chair.

Musical renditions of the folk tune “The Irish Washer Woman” or of Igor Stravinsky’s ballet score *The Rite of Spring* form historical kinds. The renditions are copied from other people, or from scores that had been transcribed from earlier renditions or had been copied from earlier scores. McDonald’s restaurants form a historical kind. There are historical, cultural causes for their being so much alike. Professors, doctors, and businessmen form historical kinds that are especially well integrated when the groups are studied in particular historical and cultural contexts. Individual professors, doctors, and businessmen are likely to act in similar ways and to have attitudes in common as a result of: similar training handed down from person to person (reproduction or copying), and/or custom (more copying), and/or natural human dispositions (compared to dog dispositions), and/or social pressures to conform to role models (copying again), and/or legal practices handed down from univocal sources. There is a reason why it may be productive to investigate, for example, “the attitudes of American doctors toward acupuncture”. These attitudes are contagious. They spread.

3.2. Eternal kinds

Compared to members of a “historical kind,” members of an “eternal kind” are like one another for a different kind of reason. They are alike because they possess a common inner nature of some sort, such as an inner molecular structure, from which the more superficial or easily observable properties of the kind’s instances flow. The inner structure results in a certain selection of surface properties, or in given selections of properties under given conditions. Popular examples of this are the various chemical elements
and compounds, along with their various forms, for example, \( \text{H}_2\text{O} \) which is ice, liquid water, or steam. Portions of water have an inner structure in common with other portions of water, which produces the same surface properties given the same temperature conditions. Strictly speaking, I suppose, gold, nitrous oxide, ice, and so forth are not members of “eternal kinds,” but rather are portions of a “stuff” (material), although samples of them are members of eternal kinds. Also, water molecules, electrons, protons, and so forth, are examples of eternal kinds. Stars, planets, comets, asteroids, and geodes are eternal kinds, not because their properties flow always from exactly the same inner nature, but because they were formed by the same natural forces, in the same sort of circumstances, out of materials that are similar in relevant ways.

Members of an eternal kind can be said to have “essences” in a very traditional sense – essences that are not nominal but real, and often discovered only through empirical investigation. The reason that the members of these kinds have many properties in common is that they have a few fundamental and/or causal properties in common, which account – given the laws of nature – for all their other properties. Eternal kinds form classes, all of whose members are alike in a variety of aspects. But they are also much more than mere classes, because they are alike in these aspects not by accident but in accordance with a causal explanation.

3.3. Individuals

The last kind of (Aristotelian) substances are individuals. Individuals have been seen in modern times to have a very different sort of unity than kinds, but there is a way in which the cement that holds a single individual together as it endures through time is quite a lot like the cement that holds a historical kind together. Ghiselin and Hull claimed that species are actually individuals, because they are held together not by a traditional essence, but rather through historical causal connections. The other side of this coin is that individuals are rather like species. A species is a “homeostatic system....amazingly well-buffered to resist change and maintain stability in the face of disturbing influences” [Eldredge and Gould (1972, p.114]. Similarly, of course, an individual animal is a “homeostatic system....amazingly well-buffered to resist change and maintain stability in the face of disturbing influences.” If a woman is tall, brown-haired, knowledgeable about electronics, and a good piano player today, it is likely, though not certain, that she will also possess these traits tomorrow. Various members of a species are like one another in part because they are, as it were, copied from one another. A physical object tends to have the same physical properties the next day as it had the previous day, because of natural conservation laws which tend not to copy, of course, but to preserve its properties from day to day. The effect, however, is much the same. The inference that an individual animal or inanimate object will probably have certain properties tomorrow because it has them today is likely to be true for the same general reason that the inference that other members of a species probably have certain properties because this member has them is likely to be true. Individual objects are things that inductive knowledge can be collected about over time; similarly, a
historical kind and, more broadly, an eternal kind are things that knowledge can be collected about over time.

I have explained why historical kinds, eternal kinds and individuals – three basic kinds of (Aristotelian) substances – are similar with respect to why it is possible to gain inductive knowledge about one part of such a cemented-together unity from other parts. The reason this is possible for each kind is that it is not merely a class, either focused or fuzzy. Because substances are not classes – not units cemented together merely by some set of common or overlapping properties – to have a concept of a certain substance is not to have a certain set of properties in mind, whether derived from paradigm cases or from exemplars. My next job then is to explain what it is to have a concept of a certain substance if not to have in mind a set of central properties. I am going to do this by explaining, first, what it is to have a concept of an Aristotelian primary substance – of an individual. Then I will generalize to other kinds of substances.

4. Concepts of individuals

The idea that there are such things as “concepts” of individuals is foreign to many psychologists and philosophers for interesting historical reasons beyond the scope of this discussion. If this use of the term “concept” bothers you, then you can interpret this as “thoughts” of individuals or “ideas” of individuals instead of “concepts of individuals.”

What is involved in being able to think of an individual? In the twentieth century, one traditional answer to this question is that to think of an individual person is to capture that individual with a description that uniquely identifies him or her. A second answer is that to think of an individual requires that you know how to identify him or her, perhaps by description, or by being able to recognize him or her, or to differentiate him or her from other individuals by perception. These views are close enough to the answer I would give myself that they will serve my purposes here. Something that they have in common, and that I am sure is correct, is the assumption that there are innumerable ways to think of the same individual. An indefinite number of individuating descriptions apply to every individual. Similarly, there are, in general, numerous ways that the same individual might be recognized by sight, by characteristic sounds, by smell (dogs are good at this) and so forth. Contrast the ways Helen Keller recognized her friends with the ways they recognized one another. Twentieth-century tradition had it, then, and I believe correctly, that there is no single or definite set of properties that one must either think of or be able to discriminate in order to have a concept, or a thought, of an individual. Nor is there some central set of properties, some or most of which one must think of or be able to recognize in order to think of a particular individual. Similarly, I will soon claim, there is no central set of properties, all or some of which one must be able to think of, recognize or discriminate in order to think of the (Aristotelian) substance *dog*, in order to learn about dogs, to understand things said about dogs, and so forth. I will come back to this a bit later.
First, we have to deal with fallibility. The ways we have of recognizing individuals are always fallible, in principle. Even presumably individuating descriptions always presuppose that there is indeed one and only one thing fitting the description, which is not guaranteed, for example, merely by a description containing superlatives. It might be true, for example, that no one is tallest or first in line, or first on the moon. More important, if you are actually to use an individuating description for purposes of recognizing an individual, you will have to recognize exemplifications of the properties mentioned in the description. But one’s capacities to recognize objective properties are always fallible, for they depend on external intervening or mediating conditions such as lighting conditions, or atmospheric conditions, or sound absorption and reflectance properties of surrounding objects, or obscuring conditions such as intervening objects that mask sounds and odors, and so forth. Nor is there an independent way of ascertaining what these mediating conditions happen to be, in a particular case. There are always possible conditions under which you would misidentify or fail to recognize even your own mother or spouse.

Having the ability to recognize an individual, then, cannot be the same thing as being infallible in recognizing him or her. For example, I have the ability to walk. It is one of my very best abilities. It does not follow that it cannot happen that I trip and fall when trying to walk. These reflections suggest that what we need here is an analysis of what it is to have an ability to do something, such as walking or recognizing your mother, that does not equate the ability with any simple sort of disposition. That analysis has been given in OCCI. But that all abilities are fallible is common sense, and I propose just to assume this here.

Tradition plus common sense suggest, then, not only that different people can have different kinds of concepts of the same individual by using quite different methods of recognition, but also that the methods any one person uses to recognize – hence to be able to think of – an individual will be fallible. Nor do these methods constitute a definition of the individual. Your mother is not defined by the way you recognize her, for example, by the look of her face and the sound of her voice. She does not have a definition, a set of properties, that make her who she is. She is not a class that happens to contain only one member.

Similarly, the species dog is a unity that different people can have quite different kinds of concepts of, by using quite different methods of recognition. Whatever method a person uses for recognizing dogs may always be fallible. Nor does the method that a person uses for recognizing dogs constitute a definition of what dogs are, even for that individual. The species dog is not just a class that happens to contain a certain number of members.

5. Concepts of substances more generally

What makes substances interesting is that there is often a great deal that can be found out and known about them. Often, they have a great many properties. Typically, numerous properties and numerous sets of these properties will each be diagnostic of the substance.
That is, each of these properties or property sets will typically only be found when the substance itself is encountered. At least this will often be so within the spatial and temporal area inhabited by the person needing to recognize a substance. Mistakes that people might have made had they lived in different places and times are not relevant to their actual abilities to recognize substances. This is why it is possible for different people to have concepts of the very same substance by very different means. Children and chemists have different ways of recognizing sugar. You and Helen Keller have different ways of recognizing nearly every secondary substance, nearly every ordinary stuff and nearly every ordinary eternal kind and historical kind.

Further, none of the ways that a person knows to diagnose the presence of a substance need to be infallible. No particular set of properties used to diagnose a substance are ever definitional of it, although in the case, especially, of eternal kinds, empirical investigation may reveal (with probability) that, in fact, some sets are always correctly diagnostic. It is always logically possible that there is some other substance that has parts of its cemented-together unity that share the very same properties as the properties one is using, with practical success, for diagnosis of a certain substance. For philosophers, this can be expressed by saying that the possibility of “Twinearth water,” certainly of “Twinearth dogs” and, indeed, of “Twinearth Mama,” indistinguishable from your mother, is never ruled out by logic alone. It takes more than a set of properties in your mind to determine a substance. It takes a certain sort of causal glue in the world, holding that substance together. But given that glue in the world, conceptual access to that glued-together unity may be had by reference to any of many of its different parts or properties.\(^1\)

6. Substances encountered through language

In talking about what is involved in having a concept of a substance, I have quietly been making an assumption that I must now bring into the foreground as a claim. I have spoken of ways of recognizing a substance, and I have said that your ability to recognize a certain substance can depend on your inhabiting a certain space-time locale, one where certain diagnostic properties mostly signify encounters with that substance rather than with others. The assumption I am making is that thinking of a substance involves the ability to recognize it, as it were, \textit{in the flesh}, not merely the ability passively to contemplate its properties. We have thoughts of substances in order to be able to collect information about substances; we pick up this information on some occasions and then apply it on other occasions. To pick up information about a substance, you must be in

\(^1\) Philosophers may detect a missing link in this analysis. The link is needed to connect the ability one has to recognize a particular substance to prior encounters with that particular substance rather than with similar substances on Twinearth or wherever. That link is supplied in the description of abilities given in OCCI. What an ability is an ability to do is determined not merely by current dispositions but by the histories of the mechanisms responsible for those dispositions.
a position to interact with the substance, or to interact with other things that interact
with the substance, are influenced by the substance, or influence it. Natural information
is transmitted in the causal order, and you have to be in the causal order with whatever
the information is about in order to receive it².

Now, if you think about that claim for a moment, you will see that it is a fairly radia-
cal one. Surely you can have a concept of the last dinosaur species on earth to go com-
pletely extinct and of the first baby to be born next year, and of any other substance
which, although you have never encountered it, you do know an identifying description.
And you have these concepts without having the slightest idea how to identify these
things in the flesh. Surely you can have a concept of molybdenum – you can think about
it and ask questions about it – without being able to identify it in the laboratory. Surely
you can have a concept of Socrates without being able to identify him in the flesh, even
if you were to be transported back to ancient Athens. Let me tackle the descriptions first,
then come back to molybdenum and Socrates, for they will prove far more interesting.

The descriptions are handled this way. That your circumstances are such that you
never get a chance to use an ability that you have does not take that ability away from
you. You will not lose the ability to swim just because they chain you to a post in the
middle of the Sahara desert for the rest of your life. If you understand the terms in any
description and know how to apply them, that is, you know how to recognize the other
objects and properties and relations mentioned in the description, and if you are right
in what the description is identifying, then you know a way to identify the substance
that the description describes. You would do so by encountering something that you can
recognize directly as fitting that description, or by coming across something else that
you recognize as carrying information telling you what fits that description. There are
many cases in which you just are not at all likely to come across any such information,
but that is irrelevant to whether you have a capacity to recognize the substance. I am
assuming here a fairly usual reading of the notion of natural information, according to
which information about the past and about the future are entirely routine kinds of
information (see footnote 2). And I am about to claim that language is a standard
medium through which natural information is transmitted, hence a standard medium
through which substances are recognized exactly as they are recognized “in the flesh”
through other media such as light and sound.

Now consider molybdenum and Socrates. It seems an obvious fact that many of our
concepts of substances have been acquired without encountering those substances
directly but only by hearing about them. Moreover, as Kripke (1972), Putnam (1975) and
Burge (1979) have observed, we often have no unique descriptions of these substances in
mind either. How then can we be said to know how to recognize them? The answer, I
claim, is that speech is just as direct a medium for the perception of objects and events

² I am using the notion of natural information somewhat, yet not quite, like the way Dretske uses it in
Knowledge and the flow of information (1980). For our purposes here, the difference probably does not mat-
ter, but a careful description of the kind of information I have in mind is in Millikan (2004, cs. 3–5) where I
call it “local information.”
and their properties as is the light reflected off objects, the smells emanating from objects, the sounds emanating from events in the environment, or the mechanical stimulations caused by objects in direct contact with one’s body. This is a thesis that requires defense, and I have defended it at length both in OCCI (Chapter 6) and in Millikan, 2004 (Chapter 9). Here, I can only throw out the rough idea, hoping that if it strikes you as dubious, you will look to these longer versions and defenses before making a final judgment.

The claim is that hearing and immediately believing a sentence about a fact or occurrence is in relevant respects just like, for example, seeing that something is the case or seeing that something has occurred and immediately believing it. There is experimental evidence that what one is told goes directly into belief unless cognitive work is done to prevent this, just as what one perceives in other ways, through other media, does. Loading the cognitive systems with other tasks, such as having simultaneously to count backwards by threes, has the effect of facilitating belief fixation regarding whatever one hears or reads [Gilbert (1993)]. Recognizing a linguistic reference to a substance is as much a way of recognizing the substance “in the flesh” as any other way of recognizing it. It is identifying it and recognizing natural information concerning it through one more medium of manifestation. Think of this medium, the speech of another person, as being like an instrument that aids perception. The lens of one’s eye is, of course, an instrument that aids perception. If one wears corrective lenses, they are another such instrument. The speech of another person is analogous to somewhat more complicated instruments of this kind. Acting like a camera, a radio, CAT scanner, or a microscope, another person who talks to you picks up information-bearing patterns from his environment, focuses them, translates them into a new medium, and beams them at you. Think of living in a language community as like being inundated in one more sea of ambient energy. Like the surrounding light, surrounding people transmit the structure of the environment to you in ways that, barring certain interferences, you have become tuned to interpret. Becoming tuned to interpret the information-bearing patterns that are common in a certain language community is coming to understand the language of that community. Similarly, a radiologist must learn to interpret the information contained on X-ray images and the auto mechanic must learn to interpret the information contained in the sounds emanating from ailing automobile engines.

The notion that understanding and believing what is said to you is just one more level of natural-sign reading, on the same level as ordinary perception, is, to many people, quite unintuitive. One reason is that what is given to you in ordinary perception is always given in some quite definite current relation to you. It is given as happening at the time you perceive it, as happening relatively nearby, and often, as bearing quite an exact spatial relation to you. This kind of information is needed to guide action, for how one can presently act on a thing always depends on its present relation to oneself. Ordinary perception is for immediate action, whereas what one learns through language is not typically used that way. Usually you are not told what exact spatial and temporal relations the objects and events being presented to you through language have to you here and now. But there are intermediate cases, for example, video recordings. It is clear enough that you perceive things happening when you watch a video, but as in the case
of language understanding, you do not perceive the spatial and temporal relation to yourself of what occurs on video.

A second reason that the comparison between ordinary perception and language comprehension is unintuitive is that ordinary perception is so much more reliable than what one hears said, at least under common circumstances. It is not easy to fool ordinary perception. To create strong perceptual illusions requires a good deal of knowledge about the perceptual mechanisms and often quite special equipment, of the kind, for example, that optometrists have in their examination rooms. This is a difference of degree, however, a mere difference in frequency, not a difference in kind. Recalling that film dubbing is currently the rule rather than the exception, what differences are there, for example, among (1) believing what you apparently see when you look through the peephole into an Ames room, (2) believing what you see when a film has been dubbed, and (3) believing what you hear someone say when it is false? In the modern world, if you want to believe only what is true, you often have to apply heavy filters to other methods of perception as well as to perception through language.

The upshot of these reflections is that we can understand how it is possible to recognize a substance through the information that language bears, and indeed, how it is possible to come to be able to recognize a substance pretty much merely by learning a word for it. This is how we manage to have concepts of Socrates and, for most of us, how we manage to have concepts of molybdenum. To have a word for a substance is to have an essential part of an ability to recognize manifestations of it that are generated in a particular language community. That, I have argued, is why it is possible for small children to learn, as Chomsky puts it, “a word an hour” between 18 months and 6 years of age [Chomsky (1995) p.15].

References


PART 3

SYNTACTIC CATEGORIES
This page intentionally left blank
Chapter 14

LEXICAL, FUNCTIONAL, Crossover, AND MULTIFUNCTIONAL CATEGORIES*

LISA DEMENA TRAVIS

McGill University

Contents

Abstract 320
1. Introduction 321
2. Categories as feature bundles 321
  2.1. The system 321
  2.2. Natural classes 322
  2.3. Unnatural classes 324
3. Categories and phrase structure 325
  3.1. Lexical and functional categories 326
  3.2. Articulation of functional categories 327
  3.3. Articulation below N and V 329
  3.4. Crossover and multifunctionality
      3.4.1. Crossover projections 330
      3.4.2. Multifunctional categories 335
4. Where do categorial distinctions reside? 336
5. Conclusions 344
References 345

* Research that provided some of the examples given in this chapter was supported by SSHRCC 410-2001-1486 and FQRSC 2002ER75657. I refer the reader to Baker (2003) and Déchaine (1993) for a much more detailed study of categories. Much of the background I give here is based on their work. I thank Malagasy consultants Rita Hanitramalala, Saholy Hanitriniaina, and Irène Rakotoanony.
Abstract

The notion of categories within the domain of syntax is fairly clearly circumscribed. We have an idea of what nouns, verbs, adjectives, and prepositions are. In this chapter, I present issues that arise, however, when specific questions are raised concerning these categories: What are the distinguishing characteristics between categories? Are categories primitives? Are category labels attached to lexical items or is their content derivable from syntactic environments? In the process of discussing these issues, noncanonical constructions such as crossover projections (e.g., gerunds) and multifunctional categories are presented. The discussion is placed within the context of natural language data that have been used to defend or refute particular claims. While the main goal of the chapter is to provide some background on the issues, I will defend a need for categorial information in lexical entries using data from St’át’imcets and Malagasy.
1. Introduction

When asked to discuss categories in the domain of syntax, one can feel secure knowing that the scope of the study is both clearly defined and quite technical, especially when compared to a similar discussion set in the context in many of the other cognitive sciences. Anyone with a Grade 4 education can rattle off at least three or four syntactic categories and define them. A noun is the name of a person, place, or thing. However, saying that categories are well known does not mean that they are well understood. In his book *Lexical Categories* (2003), Baker writes “It is ironic that the first thing one learns can be the last thing one understands.” In this chapter, I present some of the issues that arise in any investigation of the status of syntactic categories. I will discuss the use of features in creating a typology of categories [see, e.g., Chomsky (1970), Jackendoff (1977)], the distinction between functional categories and lexical categories [see, e.g., Abney (1987)], and the notion of extended projections [Grimshaw (2000)]. One of the clearest splits in categories is the one between nouns and verbs but one can ask, as I will, where this distinction comes from. Does it come from the context within which the category appears, or from something inherent to the lexical item? In order to investigate this question, I present some noncanonical instances of categories such as crossover projections (e.g., gerunds), which seem to be internally verbal but externally nominal, and multifunctional categories, which are lexical items that change function as they change environment. Finally, I use data from two languages that appear to be textbook examples of languages with no categorial distinctions, to show that, at an extremely subtle level, they display the same categorial divisions as other languages. My conclusion, then, is that categorial distinctions exist in all languages at the most basic level – the lexical entry. While this is the main goal of the chapter, a secondary goal is to develop a methodology that uses natural language data to answer questions concerning the status of categories in syntax.

2. Categories as feature bundles

Distinctions between category types were noticed by Dionysius Thrax around 100 BC with respect to inflectional affixes. While nouns (Ns) inflect for case (*rosa ‘rose.NOM,’* *rosam ‘rose.ACC’*), verbs (Vs) inflect for tense and person (*amo ‘love.PRES-1SG, amas ‘love.PRES-2SG’*). That there are distinctions has never been questioned. However, how these distinctions are captured has been, and continues to be, a concern within transformational generative grammar.

2.1. The system

Within the framework of early transformational generative grammar, these categorial distinctions were broken down further into a feature system, the goal of which was to capture generalizations of syntactic behavior across categories. Jackendoff (1977), developing the ideas of Chomsky (1970), provided the following feature system:
Categories are not primitives but feature bundles. For instance, A(djective) is shorthand for the feature bundle \([-N, +V]\). In this way, two features can be used to describe four categories as well as four natural classes, each containing two categories. Below we see some of the data that have been used to support the natural classes that this feature system describes.

2.2. Natural classes

The table in (1) predicts that A and V should form a natural class, described by the feature value \([-N, +V]\). The following two sets of data support this. In each case, a syntactic or morphological process applies to Vs and As but not to Ns and prepositions (Ps). In the first case, we find that raising of an embedded subject to a matrix subject is possible only with V/A predicates and not with P/N predicates. In the second set of data, we see that un- can be prefixed to Vs and As but not to Ps and Ns.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Feature Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>+V</td>
</tr>
<tr>
<td>N</td>
<td>-N</td>
</tr>
<tr>
<td>V</td>
<td>+V</td>
</tr>
<tr>
<td>P</td>
<td>-V</td>
</tr>
</tbody>
</table>

(2) Raising predicates

a. He seems – to be at the demonstration. √V
b. *He’s a good bet – to be at the demonstration. *N
c. *He’s on the cards – to be at the demonstration. *P
d. He’s likely – to be at the demonstration. √A

(3) Un- prefixation

a. Undo, untie, unfold, unpack, unravel √V
b. *Unfear, *unconvention *N
d. Unafraid, unaware, unfit, unkind √A

The data in (4) below show further that A and N act as a class (described by the feature value \([-N]) to the exclusion of V and P \([-N]). Both Ns and As can appear with demonstratives, while Vs and Ps cannot.

(4) Demonstratives

a. *She will [that/this] advance. *V
   cf. She will advance by [this/that] much.

---

1 The data used to both support and refute the feature system are adapted from data given in Déchaine (1993). It can be dangerous to produce such data sets without giving an account for them. Also, some questions can be raised concerning the data themselves (for example, in (2), the structure for (2b) and (2c) may not be exactly parallel to (2a) and (2d)). I will, however, let these sets stand basically as I found them in Déchaine’s work. They serve the purpose of showing the direction of the argumentation.
b. She bought [this/that] cloth. √N

c. *She was [that/this] in the water. *P
   cf. She was [that/this] much in the water.

d. It is [this/that] long. √A

The third natural class that we find confirmation of is the one described by the feature value [–V]. In (5), we see that NPs and PPs can appear in a cleft position in English, while VPs and APs cannot.

(5) Cleft
  a. *It was [go home] that they think I did. *V
  b. It was [a used car] that I wanted her to buy. √N
  c. It was [under the chair] that I think I left my coat. √P
  d. *It was [very nice to me] that you said she was. *A

Perhaps the best-known natural class is that described by the feature value [–N]. This class contains heads that are able to assign (abstract) case to their complements. It is this case-assigning feature that allows bare NP complements for Vs and Ps, as shown in (6) below.

(6) Case marking (directly followed by NP)
  a. Paul likes [NP Chinese leeks]. √V
  b. *Paul is an aficionado [NP Chinese leeks]. *N
  c. Paul is into [NP Chinese leeks]. √P
  d. *Paul is fond [NP Chinese leeks]. *A

While arguing for a system of features to describe categories, both Chomsky and Jackendoff also point to the role that phrase structure plays in the system, as well as the flexibility that a lexical item might have. Both of these issues will become important later in the discussion. They created a structural template (X-bar structure) to capture crosscategorial similarities. Further, Chomsky stresses that a lexical entry may not be tied to one category, but may only obtain a fixed categorial signature when inserted under a node in the tree. For example, while gerunds will be derived in the syntax (changing criticize to criticizing), derived nominals (such as criticism) are simply the nominal form of the verb. Jackendoff summarizes Chomsky’s proposal as follows:

Chomsky proposes that criticize and criticism form a single lexical entry, unmarked for the syntactic feature differentiating nouns from verbs. When inserted under a verb node, this entry will direct the use of the phonological form criticize; when inserted under a noun node, it will be realized as criticism. [Jackendoff (1977), p.11]

In Chomsky’s words:

Ch. 14: Lexical, Functional, Crossover, and Multifunctional Categories

Let us propose, then, as a tentative hypothesis, that a great many items appear in the lexicon with fixed selectional and strict subcategorization features, but with a choice as to the feature associated with the lexical categories noun, verb, adjective. [Chomsky (1970), p.23]
2.3. Unnatural classes

We have seen above that we can find evidence for the natural classes described by the feature system given in (1). Equally important, however, is that we should not find processes that apply to groups of elements that do not form a natural class. For example, no process should target A and P to the exclusion of V and N, or V and N to the exclusion of A and P. As Déchaine (1993) points out, we do find both of these kinds. In (7) below, we see that Ps and As can appear with measure phrases and can act as modifiers while Vs and Ns cannot. In (8), we see that PPs and APs can be directly predicated of an NP while VPs and NPs cannot.

(7) Measure phrases
   a. *She will three yards advance
      cf. She will advance by three yards.
   b. *She bought three yards cloth.
      cf. She bought three yards of cloth.
   c. They found the dead miners two miles under the surface.
   d. The tree is three feet tall.

(8) Modification
   a. *A man go to the university
   b. *A man a hero
   c. A speech beyond the bounds of decency
   d. A man older than me

On the other side, we see cases where Vs and Ns can appear in certain constructions where Ps and As cannot. In (9) below, we have a case where Ns and Vs appear to take subjects within their projections, while Ps and As do not. In (10) we see that gapping can target Vs and Ns but it cannot target Ps and As.

(9) Subjects
   a. John criticized the book.
   b. John’s criticism of the book
   c. *John against the book
   d. *John critical of the book

(10) Gapping
   a. Max plays saxophone
      and Medusa – sarrussophone.
   b. Max’s recording of Bach
      and Medusa’s – of Debussy.

2 I am not taking a stand on these data. Stowell (1983), however, has argued that all categories allow subjects as long as the environment is controlled for.
c. *Max is against the new law  
and Medusa is – the old one.

d. *Max was critical of the recording  
and Medusa was – of the performance.

Given all of these data, then, we arrive at a position where the advantages of creating a feature system do not outweigh the disadvantages, at least at the descriptive level that I have been using to present the data. I therefore turn now to some different issues concerning the level of the grammar at which categorial distinctions should be captured.

3. Categories and phrase structure

Let us return to the original observation of Dionysius Thrax – syntactic categories are distinguished by the inflectional paradigms in which they appear. Nouns appear in a case-marking paradigm while verbs appear in a tense and agreement paradigm. This is an appealing observation, but how would it be captured in current syntactic theory? In the early versions of Chomsky’s Transformational Generative Grammar, tense and agreement morphemes were attached to the verb in the syntax via a transformational rule of affix hopping shown below.

(11) Auxiliary Transformation (Affix Hopping) [Chomsky (1957), p.113]

Structural analysis: X— Af — v — Y

Structural change: X₁ — X₂ — X₃ — X₄  \rightarrow  X₁ — X₃ — X₂#— X₄

This has the effect of turning a string such as (12a) below into (12b) where each of the affixes has “hopped” over the following verbal element and marked a word boundary.

(12) a. the + man + S + have + en + be + ing + read + the + book

     b. the + man + have + S# + be + en# + read + ing# the + book

     ‘The man has been reading the book.’

A more current view of this, however, would involve head movement from a lexical category head to a functional category head that would house the inflectional material [Abney (1987), Halle and Marantz (1993)]. This involves a slightly more sophisticated phrase structure where such inflectional affixes are assumed to project their own phrasal structure within the tree.

In the rest of this section, I present issues that arise in this new view of inflection. In Sections 3.1 and 3.2, I introduce functional (as opposed to lexical) categories and discuss the expansion of the inventory of functional categories. In Section 3.3, I discuss phrase structure refinements that have been proposed for the lexical domain of the syntactic tree. With these proposals in place, in Section 3.4, I discuss how a more articulated phrase structure has the flexibility to account for crossover projections and multifunctional categories. Throughout the discussion, the question of where a categorial signature resides will be examined.
3.1. Lexical and functional categories

In (13) we see an older version of sentential structure. In (14) we see an updated version where AUX is now I (for Inflection). I is assumed to be the head of the sentence, which is now represented as the projection of I, that is, IP. In (15) we see the head movement from V (of the root ‘am- ‘love’) to I that is assumed to occur in generating the verb form *amo.\(^3\)

(13)

\[
S \rightarrow NP \rightarrow AUX \rightarrow VP
\]

(14)

\[
IP \rightarrow DP \rightarrow I' \rightarrow I \rightarrow \text{V} \rightarrow \text{VP} \rightarrow \text{I} \rightarrow \text{V} \rightarrow \text{VP} \rightarrow \text{I} \rightarrow \text{V} \rightarrow \text{DP}
\]

Within this system, it is interesting to reassess what it was exactly that Dionysius Thrax had observed. Here we see that choice of inflectional morphology is, in fact, syntactic selection (subcategorization). Just as the verb wear selects a DP (I wear a coat, *I wear

\(^3\) Just as sentences are considered to be projections of I, NPs are considered to be projections of D(eterminer). This is discussed in more detail below. In the interests of clarity, certain details have been left out, such as the articulation of I and of VP, but these will also be touched on below.
on/at a coat, *I wear) and the verb hope selects a PP (hope for rain, *hope rain), the functional category I, which contains affixal material such as tense and agreement morphemes, selects a VP (and, as we will see below, K which contains case morphemes, selects DP). Having shifted the distinguishing features of lexical categories to these functional categories, we raise other questions. Does the “V-ness” of √am- “love” reside in the lexical entry of love or is it derived from its construction with the functional category I? This is a return to the issue we encountered above where Chomsky proposes that there is a lexical item, let us call it √CRITIC-, which is neither noun nor verb before it enters the syntax. If it is inserted under an N node, it is the noun criticism: if it is inserted under a V node, it is the verb criticize. We remove the category distinction one step further here. If √CRITIC- is inserted under a node that is the complement of D it will be criticism, while if it is inserted under a node that is the complement of I, it will be criticize. In order to better understand this issue, let us take a closer look at the system of functional categories that dominate the two main lexical categories, V and N.

3.2. Articulation of functional categories

The functional domain that, in some sense, surrounds a lexical head affixally has become more and more articulated over the past 20 years. Taking a somewhat simple view of the extended projections above the semantic heads N and V, I discuss how this development of the phrase structure, along with the movement of heads, is used to account for inflectional morphology. In (15) below, we see a version of the nominal and verbal extended projections. There is a sense that there is a parallel projection in the verbal and nominal domains – the N/V projections are the θ-domain of the semantic heads, the D/I domains in some sense “place” the event/referent in time or space, and the C/K domains link the structure into the larger structure4.

(16) Nominal extended projection   Verbal extended projection

\[
\begin{align*}
\text{K} & \quad \text{K(ase)} \\
\text{D} & \quad \text{D(eterminer)} \\
\text{C} & \quad \text{C(omplementizer)} \\
\text{I} & \quad \text{I(ndeflection)} \\
\text{KP} & \quad \text{CP} \\
\text{DP} & \quad \text{IP} \\
\text{N} & \quad \text{N} \\
\text{…} & \quad \text{…} \\
\text{VP} & \quad \text{…} \\
\end{align*}
\]

\[\text{NOM, ACC, etc.} \quad \text{the, a, this, that, …} \quad \text{that, for, …} \quad \text{PRES, PAST, etc.; + 1PL, 1SG, 1PL, 2SG, etc.}\]

4 Ken Hale discussed this sort of parallelism in his MIT classes in the 1980s. He never put these ideas into print as far as I know.
Sometimes these heads are realized by separate morphemes (words), as in the Malagasy example below, where the inflectional-type elements within the extended projection of the noun remain separate from the noun.

(17) **Malagasy**  an’ ny boky “the book-ACC”

```
             KP
            /   \
           K     DP
           |      |
          an     |
            |
            D     NP
            |
           ny    |
            |
            N     ...
            |
            boky
```

In German, there is head movement of the Determiner into the K(ase) head encoding both definiteness and case distinctions on one word.

(18) **German**  den Mann

```
             KP
            /   \
           K     DP
           /   \  |
          D     D  |
         / \  /   |
        en  d    |
         \   \   |
          N     ...
          |
          Mann
```

Taking one more example, in Romanian we can see movement of a head, in this case the noun, across two lexical items (acesti ‘these’ and trei ‘three’) to the Determiner position, where it picks up an affix that encodes definiteness.

(19) **Romanian**  [Ungureanu (2002)]

a. acesti trei copii
   these three children

b. copii-i acestia trei
   children-DET these three

---

5 Malagasy is a Western Malayo–Polynesian language spoken in Madagascar.
In all of the cases we have seen above, nouns appear in a certain environment – one that can be characterized by either the lexical items or the morphological affixes. Our concern is how the nominal characteristics of the semantic head are determined. Must the lexical item be listed in the lexicon with a categorial signature, or can this signature be derived from the syntactic environment of the lexical item? In the next section, we turn to a further complication created by recent developments in phrase structure.

3.3. Articulation below N and V

Just as there was an articulation of phrase structure above the categories N and V with the addition of various functional categories, there has also been an articulation of the lexical categories themselves. While in English, verbs such as thin in *The chef thinned the soup* appear to be monomorphemic, in other languages such predicates are bimorphemic, containing a lexical causative morpheme. There is a developing consensus that even the monomorphemic verbs of English contain hidden structure. Various researchers [(e.g., Larson (1988), Hale and Keyser (1993), Baker (2003)) assume what is called a VP shell analysis of the verb phrase. The assumption is that the VP contains (at least) two layers of verbal structure. One such analysis is given in (21) below. In this analysis, the verb thin has, as its root, the adjective thin. Within the VP there is the lower V that would encode a meaning something like BECOME, while a higher “little v” encodes a meaning something like CAUSE. A close (but not identical) paraphrase would be *The chef caused the soup to become thin*.

(21) a. The chef thinned the soup.

b.
While in English, such complex syntax for an apparently simple lexical item seems unnecessary, in other languages the complexity of the lexical item is much more evident. In Malagasy, for instance, it is much clearer that the majority of verbs are derived from nonverbal roots. Some examples are given below where verbs are generated from adjectives and nouns with the addition of the verbal prefix *man*-

(22) Malagasy

ADJECTIVES:

- *lany* ‘used up’
- *madio* ‘clean’
- *mandany* ‘to use up’
- *manadio* ‘to clean’

NOUNS:

- *lavaka* ‘hole’
- *doka* ‘flattery’
- *mandavaka* ‘to pierce’
- *mandoka* ‘to flatter’

With the articulation of phrase structure both of the functional domain that appears in the tree above the lexical category head and of the lexical category itself, complications are introduced into the system that raise questions concerning the status of syntactic categories. Before returning to the central question of the chapter, however, we will look at two noncanonical instantiations of categories – crossover projections and multifunctional categories. Accounts for both of these phenomena benefit from the fine-tuning of the phrase structure that has just been discussed.

3.4. Crossover and multifunctionality

Crossover projections and multifunctional categories challenge the notion of syntactic categories in different ways. Crossover projections are those projections that appear to be of one sort internal to the projection and of another sort external to the projection. For example, gerunds appear to be verbal with respect to the distribution of many elements within the projection, but the projection itself appears in nominal environments. Multifunctional elements are lexical items that appear to behave as different categories in different environments. The lexical item *that*, for example, is able to appear in constructions with NPs, in which case we consider it to be a determiner (*that child*); it can also appear in constructions with an IP (sentence), in which case we consider it to be a complementizer (*that the earth is round*). We look at crossover projections and multifunctional categories in turn.

3.4.1. Crossover projections

English gerunds provide an interesting testing ground for theories of categories for a variety of reasons [see, e.g., Abney (1987), Valois (1991)]. First, as mentioned above, gerunds have characteristics of nouns and of verbs. Further, there are different types of gerunds that differ in how nominal or how verbal they are. Let us begin by noting that
there are three types of syntactic realizations of gerunds, given in (23) below. \textit{Acc-\textit{ing}} is considered to be the most verbal and least nominal, and -\textit{ing of} the least verbal and the most nominal. The characteristics to note are the form that the subject of the gerund takes (\textit{Mary/them} vs. \textit{Mary's/their}) and whether or not the object of the gerund is preceded by the preposition \textit{of}.

(23) a. \textit{Acc-\textit{ing}}
   I approve of \textit{Mary/them} singing the song.

   b. \textit{Poss-\textit{ing}}
   I approve of \textit{Mary's/their} singing the song.

   c. \textit{\textit{-ing} of}
   I approve of \textit{Mary's/their} singing \textit{of} the song.

A fourth type of gerund can be detected when meaning is controlled for. The syntactic realization of -\textit{ing of} can have two different meanings, as the data in (24) show. Abney discusses this and labels the two meanings: the \textit{ACT} meaning and the \textit{FACT} meaning. \textit{Acc-\textit{ing}} and \textit{Poss-\textit{ing}} can only have the \textit{FACT} reading while -\textit{ing of} gerunds can have both the \textit{ACT} and the \textit{FACT} readings.

(24) \textit{ACT vs. FACT} [Abney (1987)] (adds fourth-type)

\begin{itemize}
  \item \textit{Acc-\textit{ing}}
    \begin{itemize}
      \item a. John fixing the sink was surprising. \textit{FACT} \checkmark
      \item b.* John fixing the sink was skillful. \textit{ACT} \times
    \end{itemize}

  \item \textit{Poss-\textit{ing}}
    \begin{itemize}
      \item c. John’s fixing the sink was surprising. \textit{FACT} \checkmark
      \item d.* John’s fixing the sink was skillful. \textit{ACT} \times
    \end{itemize}

  \item \textit{-\textit{ing} of}
    \begin{itemize}
      \item a. John’s fixing of the sink was surprising. \textit{FACT} \checkmark
      \item b. John’s fixing of the sink was skillful. \textit{ACT} \checkmark
    \end{itemize}

We can see that the different meanings have different syntax, as only the \textit{ACT} reading allows gapping of the gerund.

(25) \textit{\textit{-ing of}} (A)

\begin{itemize}
  \item a. John’s fixing of the sink was skillful and Bill’s [e] was more so.

  \item \textit{-\textit{ing of}} (F)
    \begin{itemize}
      \item b.* John’s fixing of the sink was surprising and Bill’s [e] was more so.
    \end{itemize}

  \item \textit{Poss-\textit{ing}}
    \begin{itemize}
      \item c. * John’s fixing the sink was surprising and Bill’s [e] was more so.
    \end{itemize}

  \item \textit{Acc-\textit{ing}}
    \begin{itemize}
      \item d. * John fixing the sink was surprising and Bill [e] was more so.
    \end{itemize}

\end{itemize}
The syntactic behavior of the four different types of gerunds is summarized in the table below.

(26) Four types of gerunds with tests

<table>
<thead>
<tr>
<th></th>
<th>Gapping</th>
<th>of</th>
<th>Poss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc-\text{\textit{ing}}</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Poss-\text{\textit{ing}}</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>-\text{\textit{ing}} of (F)</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>-\text{\textit{ing}} of (A)</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

In many accounts of gerunds [Abney (1987), Valois (1991)], these four levels of “nouniness” have been accounted for by attaching the \textit{-ing} affix to different places in the extended projection of the verb. The nominal characteristics of the affix (indicated by $+N$ here) will cause the projection to crossover from the verbal extended projection to the nominal extended projection at the point at which the affix has been inserted into the phrase structure. I have sketched a possibility below (and have added an additional functional category E(vent) from my own work). The details are not as important as the general type of solution that is proposed$^6$.

(27)

It is not only English that shows this type of behavior with nominal morphology. Malagasy \textit{f}-nominals also exhibit crossover within the projection and multifunctionality of the \textit{f}-morpheme$^7$. There are three types of \textit{f}-nominals – complex event nominals, result nominals, and object nominals. Like gerunds, these nominals vary in their syntactic representation and in their meaning, as evidenced by the type of predicates that they can appear in construction with. The main distinguishing factor of the syntactic representation is the type of thematic arguments that they appear with. As shown

---

$^6$ While gerunds are introduced here as a type of crossover projection, the \textit{-ing} morpheme also introduces the notion of multifunctionality [see Lefebvre (1998)], as it can appear in different positions on the tree.

$^7$ Data are taken from Hanitriniaina and Travis (1998).
below, complex event nominals appear with an Agent and a Theme, result nominals with just a Theme, and object nominals with no thematic arguments at all. To force the different readings, different predicates are used – *maharitra ela* “last a long time” to distinguish complex event nominals from result nominals, *vita tsara* “well done” to distinguish result nominals from object nominals, and a locative PP for object nominals.

(28) Complex event nominal (*Agent and Theme*)

- a. Maharitra ela ny fanoritan-dRabe ny saritany
  \[ \text{PRES-last long-time the sketching-Rabe the map-his} \]
  ‘Rabe’s sketching of the map takes a long time.’

- b. Vita tsara ny fanoritan-dRabe ny saritany
  \[ \text{done good the sketching-Rabe the map-his} \]
  ‘Rabe’s sketching of his map is well done.’

(29) Result nominal (*Theme, no Agent*)

- a. * Maharitra ela ny fanoritana ny saritany
  \[ \text{PRES-last long-time the sketching the map-his} \]
  ‘The sketching result of his map takes a long time.’

- b. Vita tsara ny fanoritana ny saritany
  \[ \text{done good the sketching the map-his} \]
  ‘The sketching result of his map was well done.’

(30) Object nominal (no Agent, no Theme)

- a. ny fanoritana ‘the instrument used for sketching’

- b. Eo ambonin’ny latabatra n fanoritana
  \[ \text{there above-the table the sketching.pencil} \]
  ‘The instrument for sketching is on the table.’

There is another interpretation for the genitive DP (the DP that immediately follows the head noun) other than that of Agent. When such a genitive DP appears with the object nominal, it is interpreted not as the Agent but as the Possessor. Further, the meaning of the predicate *vita tsara* when used with an object nominal will change from “well done” to “well made.” These two facts are exemplified below.

(31) a. Eo ambonin’ny latabatra ny fanoritan-dRabe
  \[ \text{there above-the table the sketching.pencil-Rakoto} \]
  ‘Rakoto’s instrument for sketching is on the table.’

- b. Vita tsara ny fanoritana.
  \[ \text{done good the sketching.pencil} \]
  ‘The instrument is well made.’
  ≠ ‘The sketching was well done.’

---

8 See Grimshaw (1990) for a discussion of event nominals and for the use of predicates to distinguish nominal types.
To summarize what we have just seen, there are three types of *f*-nominals in Malagasy. The actual form of the head nominal is the same but what differs is the event structure and the argument structure. With the complex event type of nominal, there is a full-argument structure. In the resulting event nominal, no Agent can be realized. In the object nominal, no direct arguments can be realized. The results of the predicate tests are shown in the table below.

(32) **Three** types of *f*-nominal and tests

<table>
<thead>
<tr>
<th>Event</th>
<th>Well-done</th>
<th>Long time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Object</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

As in the case of the English gerunds, we can provide an account of these different uses of the *f*-nominal by positing a multifunctional *f*-morpheme that attaches at different points in the syntactic structure, triggering a crossover from the verbal extended projection to the nominal extended projection. This crossover will be triggered by a categorial feature attached to the *f*-morpheme.

(33) We have just seen two cases of crossover projections. The nominalizing *-ing* in English and the nominalizing *f*- in Malagasy always serve the function of changing a verbal projection into a nominal projection. In this way, in every form of their use, they contain the categorial feature [N]. They serve more than one function, though, in that they create nominals of different types depending on where they are placed on the tree. In this sense, they are not only crossover morphemes, they also serve a variety of functions. We turn now to the issue of multifunctionality.
3.4.2. Multifunctional categories

The type of multifunctionality of lexical items (independent words or bound morphemes) that we have just seen has been discussed by, among others, Lefebvre (1998), Roberts and Roussou (2003), and Roberts within the generative framework. In the two examples we have seen so far, the category in each function has remained the same. In fact, the category itself is the sum of what the affix means. We have also seen at the end of Section 2.2 the suggestion that roots in English can be category-less (like criticize/criticism), where the category is only determined once the lexical item has been inserted into a syntactic structure. In the case of criticism/criticize, the assumption is that only the category changes but the argument structure remains the same. In this section, I briefly introduce a case where the same lexical item can appear to have a different categorial signature as well as different argument structure depending on the details of the syntactic realization.

Déchaine (1993) discusses an element in Oweré Igbo that can appear in three different forms with three different functions\(^9\). It can appear as a separate lexical item with a derived low tone. In this form, it has the function of an auxiliary and the meaning of future. It can appear as a root that takes suffixes, in which case it functions as the main verb meaning “to go.” Finally, it can appear as a suffix itself, in which case it functions like an aspectual morpheme and has the meaning of a progressive. This is summed up in (34) below, followed by examples of the progressive (35a) and future uses (35b).

(34) a. main verb ‘go’ gà- (derived low tone)
   b. progressive suffixal -ga (toneless, as an affix)
   c. future auxiliary gà (derived low tone)

(35) a. O rí- gà ri à
   3SG eat PROG food this
   ‘S/he is eating this food.’

   b. O gà e-rí- ri à
   3SG FUT ?-eat food this
   ‘S/he will eat this food.’

This multifunctional element in Oweré Igbo varies with respect to argument structure. The main verb takes arguments while the aspectual marker and the tense marker do not. Also, there is a clear shift in category. What remains constant is some slightly abstract meaning of “on the way to some point.” In the case of the future marker, the point is the beginning of the event. In the case of the aspectual marker, the point is the end point of the event, indicating that that event is still in progress. For the main verb, the predicate names the content of the action itself.

There is a view of this type of multifunctionality [see, e.g., Lefebvre (1998)] whereby lexical elements can appear in the lexicon unspecified for such things as argument

\(^9\) Oweré Igbo is a Niger-Congo language spoken in Nigeria.
structure and category signature. This view is a bit stronger than the criticize/criticism case we saw above since here even the argument structure is left underspecified. In both cases, however, the underspecified material can be supplied by the syntactic environment\(^{10}\). What we can learn from multifunctional elements is that the categorial content of a lexical element is sometimes provided by the syntactic structure.

We can now ask the question of whether the categorial information is ever linked to the lexical item. In current syntactic theory within the Chomskian paradigm, this question has been given two diametrically opposed answers. We turn to the current stage of syntactic structure within this paradigm and then tackle the question of where categorial information resides.

4. Where do categorial distinctions reside?

In the Minimalist Program [see, e.g., Chomsky (1995)], Chomsky outlines a direction of research, the goal of which is to create a grammar of minimal complexity. For example, the number of syntactic levels is reduced from four (D-structure, S-structure, PF, LF) to two (PF, LF). These two levels – PF, which is the articulatory–perceptual interface, and LF, which is the logical–intensional interface – represent the “virtual conceptual necessity” of grammar as it encodes sound–meaning pairs. Within the context of the Minimalist Program, phrase structure is also pared down. In the Bare Phrase Structure of the Minimalist Program, technically, categories do not exist independently of the lexical items that carry the labels. In older versions of the Chomskian paradigm, lexical insertion involved a matching of the category (and subcategory) of a lexical item with a category in a preconstructed tree. In the Minimalist Program, lexical items are MERGED with one another (put in a sister relationship), and the content of the tree is determined entirely by the features of the lexical items themselves. In Chomsky’s words:

In particular, there is no way to project from a lexical item \(\alpha\) a subelement \(H(\alpha)\) consisting of the category of \(\alpha\) and whatever else enters into further computations… we thus dispense with such structure as [(36a)] … in place of [(36a)a] we have [(36b)]


(36)

\[
\begin{align*}
\text{a.} & \quad \text{DP} \\
& \quad \text{NP} \\
& \quad \text{the} \\
& \quad \text{N+} \\
& \quad \text{book} \\
\text{b.} & \quad \text{the} \\
& \quad \text{book}
\end{align*}
\]

\(^{10}\) See Hale and Keyser (1993, 2002) and Erteschik-Shir and Rapoport (1997) for some views of how argument structure can be encoded in syntactic structure.
In (36b), two lexical items, the and book, are merged. The mother node that dominates both items, like each of the items itself, is a set of features. Selectional properties will determine which of the lexical items will percolate its features to the dominating node (i.e., which is the head of the construction); in this case, the head is the determiner the. Structure (36b) represents the fact that all of the features of the lexical item percolate, not just the category. The structure cannot add information either. In other words, the lexical item cannot be underspecified with respect to category and have its category supplied by the tree itself.

While Chomsky’s intent is not to investigate the details of a lexical entry—“I have little to say about the lexicon here” [Chomsky (1995), p. 235]—he does specifically mention that lexical items appear with a categorial signature:

For the word book, it seems the optimal coding should include a phonological matrix of the familiar kind expressing exactly what is not predictable, and a comparable representation of semantic properties, about which much less is known. And it should include the formal features of book insofar as they are unpredictable from other properties of the lexical entry: perhaps its categorial feature N, and no others.


This view of the lexicon and of category labels in particular raises questions about the status of, say, √DESTROY. Does such a lexical item appear in the lexicon with a categorial feature? Are there separate lexical items for [N, destruction] and [V, destroy]? Now we turn to a different view of categories that remains within the Minimalist Program but which pushes Chomsky’s (1970) notion of category-less roots even further.

Marantz (1997) develops a view of Bare Phrase structure that allows MERGE to manipulate category-less roots. Since category labels cannot exist independently of a lexical item, i.e., the tree structure itself cannot add information, the category-less root must be disambiguated (between destroy and destruction) by some other means. The way this is done is by having subsequent instances of MERGE disambiguate the categorial status. As shown in (37) below, the projection of √DESTROY remains category-less, but when that projection MERGES with a determiner (or rather a lexical item which contains a determiner feature), the resulting structure will have the characteristics of a DP and the form of the lower head will, in this environment, be spelled out as destruction.

(37) The destruction of the city [Marantz (1997), ex. 16]

![Diagram]

Ch. 14: Lexical, Functional, Crossover, and Multifunctional Categories
On the other hand, if the category-less projection of the category-less root MERGES with a lexical entry with a verbal feature, the result will be the verb \textit{destroy}^{11}.

\begin{equation}
\text{John destroyed the city}
\end{equation}

While this is appealing and manages to both capture the flexibility observed in Chomsky (1970) and adhere to the restrictions imposed by Chomsky (1995), it raises several questions. One concerns the productivity of such a process. Which lexical items are category-less? Marantz takes a strong stand and proposes that basically all lexical categories (nouns, verb, adjectives) are formed in a similar fashion from category-less roots. This claim predicts that we should find productive triplets (noun, adjective, verb) of every category-less root. This is clearly not the case, however. For example, while √\textit{CUP} makes a good noun and a good verb, this is not the case with √\textit{SAUCER}. Marantz notes this problem but suggests that while, in principle, all should be possible (i.e., √\textit{SAUCER} should be found in a verbal form), only some forms become familiar through use. He writes:

\begin{quote}
The interaction of root semantics and the semantics of the heads that create nouns, verbs and adjectives determines how good a combination of a root and such a head will be. So, while “cat” as a noun is fine, as is “cat” as an adjective (“catty”), “cat” as a verb has no obvious meaning/use, although it can be given fine meanings contextually. (“Meowing and scratching in imitation of his pet feline, Fred catted around the house for hours.”) 
\end{quote}


One thing that makes Marantz’s view appealing is that there are languages, such as St’át’imcets (Lilloet Salish)\textsuperscript{12}, which will be discussed below, where lexical items really do appear to be underspecified as to which category they belong to. For languages such as these, a theory of category-less roots appears quite appropriate. Before turning to St’át’imcets, however, let us look at an alternative view being proposed currently.

Baker (2003), also working within the Minimalist Program, presents a diametrically opposed view. For him, lexical roots are crucially not category-less. While proposing that the information necessary in the lexical entry includes categorial information is not novel, the way that Baker has the categorial signature play out in the syntax is novel.

\begin{itemize}
\item In English, this verbal category, called here v-1, is null. This is, however, an example of the articulation of VP that we saw in Section 3.3.
\item St’át’imcets is a Salish language spoken in British Columbia.
\end{itemize}
The details of his analysis are not crucial to the discussion here, but it is important to note that his aim is to account for syntactic differences between categories such as those discussed in Section 2.2 through inherent syntactic differences in the lexical entries themselves.

Having set up these two very different views of categorial status, let us turn back to the data. In St’át’imcets, lexical categories do seem to be underspecified for lexical category. As we can see below, roots can appear in a verbal context, as in (39) below, or a nominal context, as in (40) below [data from Demirdache and Matthewson (1995)]. It is important to note the role that morphology and syntactic environment play in the disambiguation of these forms. In (39), the root appears with agreement morphology that makes it look verbal. In (40), the root appears with a determiner and the appropriate nominal morphology, making it look nominal.

(39) a. qwatsáts-kacw ‘You left/you leave’
   leave-2SG.SUBJ
   
   b. smúlhats-kacw ‘You are a woman’
   woman-2SG.SUBJ
   
   c. xzúm-lhkacw ‘You are big.’
   big-2SG.SUBJ

(40) a. qwatsáts ti smúlhats-a ‘The woman left.’
   leave DET woman-DET
   
   b. smúlhats ti qwatsáts-a ‘The one who left is a woman.’
   woman DET leave-DET
   
   c. qwatsáts ti xzúm-a ‘The big one left.’
   leave DET big-DET

The facts of St’át’imcets can easily be represented through an account where roots are category-less and can appear in a verbal environment which is typified by being in a projection that is the complement of the I(nflectional) node (which would house the agreement morphology), as in (41a). The nominal version would be created through MERGE with the nominal functional category, Determiner, as in (41b).

(41)

```
a. IP       b. DP
   I         D
   √         √
```

St’át’imcets, then, appears to be the kind of language for which a theory such as Marantz’s works best. However, as Demirdache and Matthewson (1995) show, in spite of a vast number of examples that point to a category-less lexicon, there are several
constructions which crucially are sensitive to some presyntactic category label. An example of one such construction, the relative clause, is given below. We have already seen above that roots that would be given an adjectival translation in English (e.g., *big*) can serve as nouns and verbs in the role of the subject of a sentence or the predicate of a clause, respectively. Examples (42) and (43) show, however, that these adjective-like roots cannot serve as heads of a certain type of relative clause. Example (42) introduces the structure using roots with core nominal meanings, while (43) shows that only these nominal roots, and not adjectival roots, can be found in this construction.

(42) a. *ats’x-en-lhkan [ti qwatsáts-a ʔsqayew] see-TR-1SG.SUBJ DET leave-DET man
     ‘I saw the man who left.’

     b. *ats’x-en-lhkan [ti xzúm-a ʔspzúza ʔ7] see-TR-1SG.SUBJ DET big-DET bird
     ‘I saw the bird who is big.’

(43) a. *ats’x-en-lhkan [ti xzúm-a ʔtseqwtsíqw] see-TR-1SG.SUBJ DET big-DET red
     ‘I saw the red one who is big.’

Again, this is particularly surprising in a language where, in garden variety constructions, there seems to be little or no distinction between category types and little reason to encode any differences in the grammar.

Malagasy is another language that allows a certain amount of categorial ambiguity, though less than we have seen for St’át’îmcets. Below we see examples where a variety of elements can act as predicatives.

(44) a. Antira io olona io.
     old DEM person DEM
     ‘This person is old.’

     b. *matory io olona io.
     sleep DEM person DEM
     ‘This person is sleeping.’

     c. Mpianatra io olona io.
     student DEM person DEM
     ‘This person is a student.’

Furthermore, all of these forms can also appear in a nominal context, much like the cases we saw in St’át’îmcets.¹³

¹³ As mentioned in the text, this flexibility is not as robust as in St’át’îmcets. One has to work a bit to get examples that work, and my consultant finds that the example given in (45b) requires a context to make it less awkward.
Again we have a language where there seem to be underdetermination of category labels. But, again as in St'át'imcets, there are constructions which make a distinction. In Malagasy, the construction is quite peripheral to the grammar which, I believe, makes it even more interesting. The core constructions are the ones where categories seem quite flexible, while the contrasts show up in quite subtle ways.

Keenan and Ralalaoherivony (2000) introduce a construction they label Agent Phrase Raising. An example is given in (46) below. The structure is quite complex and the analysis not obvious, but for our purposes, only a basic knowledge of the facts is necessary. In the nonraised example (46a), the subject consists of a noun, raharaha “work,” modified by a relative clause, sahaniko “faced by me,” which contains an Agent –ko “me.” In the raised construction (46b), the Agent has raised not only out of the relative clause but out of the nominal and now stands alone as the subject aho “I.” The remaining material is found within the VP [data from Keenan and Ralalaoherivony (2000)]. An additional example from their paper is given in (47). Subjects are underlined and agents put in bold to clarify the constructions.

(46) a. \[ vp \text{ Maro} \] \[ ny \text{ raharaha} \] \[ sahaniko \] much DET work TT-confront-1SG
   \‘The work faced by me is great.’

b. \[ vp \text{ Maro} \] \[ raharaha \] \[ sahanina \] aho
   much work TT-confront 1SG
   Literally: ‘I am much work confronted.’

(47) a. Tsy tonga \[ ny \text{ vahiny} \] \[ nasai\text{ko} \] NEG arrive DET guest PST.TT.invite.1SG
   ‘The guests that I invited have not arrived.’

b. Tsy tonga vahiny nasaina aho
   NEG arrive guest PST.TT.invite 1SG
   ‘I am in the state of having the guests not show up.’

While for most speakers these expressions are quite restricted in use, for some speakers the construction is fairly productive and this latter group shows an interesting
asymmetry in the examples\textsuperscript{14}. It appears that the category of the main predicate is important in determining the detailed behavior of the construction. In the end, there are two classes of construction – the adjectival construction and the verbal construction, which I will call the “true raising” construction\textsuperscript{15}. More examples of each type are given below.

(48) Adjective roots (stative): (like maro ‘many’ in (46)
\begin{itemize}
  \item a. Tsy lany ny zavatra irin’ny olombelona
      \begin{center}
        NEG exhausted DET things TT.desire DET humans
      \end{center}
    ‘The things desired by humankind are unending.’
  \item b. Tsy lany zavatra irina ny olombelona
      \begin{center}
        NEG exhausted things TT.desire DET humans
      \end{center}
    ‘Humankind has limitless need of things.’
\end{itemize}

(49) Verb roots (eventive): (like tonga ‘arrive’ in (47)
\begin{itemize}
  \item a. Tsy lasa ny vahiny nasaiko
      \begin{center}
        NEG leave DET guest PST.TT.invite.1SG
      \end{center}
    ‘The guests that I invited haven’t left.’
  \item b. Tsy lasa vahiny nasaina aho
      \begin{center}
        NEG leave guest PST.TT.invite 1SG
      \end{center}
    ‘The guests that I invited haven’t left.’
  \item c. Ho avy ny mpianatra tiaiko
      \begin{center}
        FUT come DET student TT.like.1SG
      \end{center}
    ‘The students that I like will come.’
  \item d. Ho avy mpianatra tiana aho
      \begin{center}
        FUT come student TT.like 1SG
      \end{center}
    ‘The students that I like will come.’
\end{itemize}

There are three ways in which these two forms of the construction can be distinguished. In Travis (2001), I argue that the adjectival form of the construction is base-generated and not created through movement. As such, it often has an idiosyncratic interpretation rather than a meaning that is derived compositionally (see the translations above) and in some cases can only appear in the “raised” form, as shown in (50) below.

\textsuperscript{14} The speaker who first brought the asymmetry to my attention speaks the Betsileo dialect of Fianarantsoa, which is considered to be more conservative than the main dialect, Merina. A Merina speaker from Moramanga, however, also provided similar contrasts. The bulk of Merina speakers, however, do not use this construction productively and therefore do not have what I call “true raising.”

\textsuperscript{15} I discuss this construction in Travis (2001).
Further, only the adjectival form of the construction can undergo causativization, as shown in (51) and (52).

(51) a. Mahamaro ny raharaha sahaniko
PRES.CAUS.much DET work TT-confront-1SG
ny fidiran’ny mpianatra
DET return’DET students
‘The return of the students causes there to be much work confronted by me.’

b. *Mahamaro raharaha sahanina
PRES.CAUS.much work TT-confront
ahy ny fidiran’ny mpianatra
1SG DET return’DET students
‘The return of the students causes me to have much work to be confronted.’

(52) a. Mahatonga ny vahiny nasaiko
PRES.CAUS.arrive DET guest PST.TT.invite-1SG
ny fampisehona
DET exposition
‘The exposition caused the guests invited by me to arrive.’

b. *Mahatonga vahiny nasaina ahy
PRES.CAUS.arrive guest PST.TT.invite 1SG
ny fampisehona
DET exposition

The last distinction is that the relative clause is optional in the adjectival form and obligatory in the verbal form.

(53) a. [vp Maro raharaha (sahanina)] aho
much work TT-confront 1SG
‘I have much work.’

b. [vp Tsy tonga vahiny *(nasaina) aho
NEG arrive guest PST.TT.invite 1SG
Intended reading: My guests haven’t arrived.
These characteristics are summed up in the following table\(^ {16} \).

\[
\begin{array}{|c|c|}
\hline
\text{Characteristics} & \text{Adjective} & \text{Verb} \\
\hline
\text{Idiosyncratic meaning} & \checkmark & \ast \\
\text{Causative} & \checkmark & \ast \\
\text{Optional relative clause} & \checkmark & \ast \\
\hline
\end{array}
\]

Whatever the eventual account of these constructions is, it is clear from the data given that there has to be a sensitivity to the category of the root. It is also clear that the category cannot be determined by the syntactic environment, as suggested by Marantz, since the syntactic environment is identical in both constructions.

Further, as pointed out earlier, the constructions in both St’át’imcets and Malagasy that require categorial sensitivity are not the core or most frequent constructions in these languages. This suggests that categorial distinctions must be attached to lexical items from the beginning and are simply obscured in the core constructions. They do, however, become visible at some of the edges of the grammar. As a last point, it is interesting that these distinctions are required in the types of languages that most closely match the sorts of categorial ambiguities that Marantz’s representations predict. These are languages that, for the most part, show great flexibility in the use of categories. If categorial distinctions are needed even in these languages, it raises questions about whether we should try to push a category-less grammar for other, less obvious, languages like English.

5. Conclusions

In conclusion, categories within linguistic theory and syntactic theory in particular, have a very specific connotation – syntactic categories such as N(oun), V(erb), and recently D(eterminer) and I(nflection). The concerns that syntacticians have with categories are quite technical and involve the manner of encoding categories within a formal system. Two relevant questions are:

(i) Are syntactic categories primitives?
(ii) Where and how are they encoded in the grammar?

The answers to these questions are found through the careful study of natural language data and generalizations of the sort described briefly above. As an answer to the second question, I believe that the data given suggest that lexical entries need to have the option

\(^{16} \text{The point that I was making in Travis (2001) was that roots with the true raising construction could only do this with verbal predicates, and it is true raising that forces a compositional meaning, which is ruled out in causative constructions, and which requires the presence of a relative clause.}\)
of containing categorial information that can determine the syntactic configurations in which a lexical item may appear. Certain lexical items may, however, remain underspecified, which would create the multifunctional elements discussed. A careful study of such underspecification may provide an answer to the first question since underspecification may involve a complex system of features. The nature of categories in syntax, while apparently fundamental, nevertheless remains an area where careful crosslinguistic work still needs to be done.

References

Chapter 15

ISOLATING-MONOCATEGORIAL-ASSOCIATIONAL LANGUAGE

DAVID GIL

Max Planck Institute for Evolutionary Anthropology

Contents

Abstract 348

1. Introduction 349

2. What IMA Language is Like 349

  2.1. Isolating 350

  2.2. Monocategorial 350

  2.3. Associational 351

3. Where IMA Language Is Found 354

  3.1. Semiotics 354

  3.2. Phylogeny 356

  3.3. Ontogeny 358

4. Typology 359

  4.1. Riau Indonesian: overview 360

  4.2. Riau Indonesian: analysis 364

    4.2.1. Syntax 364

    4.2.2. Semantics 366

    4.2.3. Further analysis 368

  4.3. Riau Indonesian: A Relative IMA Language 374

5. Cognition 375

Acknowledgments 377

References 377
Abstract

Isolating-Monocategorial-Associational (IMA) Language is language with the following three properties: (a) **morphologically isolating**, without word-internal morphological structure; (b) **syntactically monocategorial**, without distinct syntactic categories; and (c) **semantically associational**, without distinct construction-specific semantic rules, the compositional semantics relying instead on the Association Operator, which simply states that the meaning of a composite expression is associated with the meanings of its constituents in an underspecified fashion. IMA Language is present in the following five domains: (a) **semiotics**: some artificial languages are IMA Language; (b) **phylogeny**: at some stage in evolution, early language was IMA Language; (c) **ontogeny**: at some stage in acquisition, early child language is IMA Language; (d) **typology**: some languages are closer than others to IMA Language; and (e) **cognition**: IMA Language is a feature of general human cognition. The main part of this chapter is devoted to showing how one particular natural language, Riau Indonesian, comes close to constituting IMA Language.
1. Introduction

Imagine a hypothetical language, either natural or artificial, with the following three properties:

(1) (a) **Morphologically Isolating**
    No word-internal morphological structure;

(b) **Syntactically Monocategorial**
    No distinct syntactic categories;

(c) **Semantically Associational**
    No distinct construction-specific rules of semantic interpretation
    (instead, compositional semantics relies exclusively on the Association
    Operator, defined in (2) below).

Such a language may be described as **Isolating-Monocategorial-Associational**, or for short, **IMA**.

Does IMA Language exist? Obviously, English is neither an IMA Language, nor, to the best of my knowledge, has any other natural language been proposed to be in complete possession of the three defining properties in (1) above. Nevertheless, in this chapter, it is argued that the notion of IMA Language is relevant to a variety of domains: semiotics, phylogeny, ontogeny, typology, and cognition. In particular, the main part of this chapter is devoted to showing that although no natural language is IMA Language *per se*, some languages may indeed come much closer to exhibiting the three properties in (1) than is commonly supposed.

2. What IMA Language is Like

The three defining properties of IMA Language pertain to three different linguistic domains: morphology, syntax, and semantics. Logically, they are thus independent of each other. Accordingly, one may imagine various other kinds of hypothetical languages with different subsets of the three properties, for example, a language that is isolating but not monocategorial or associational.

The defining properties of IMA Language represent the limiting points of maximal simplicity within each of the three domains: morphology, syntax, and semantics. Hence, for each domain, one may imagine languages approaching these end points along a scale of decreasing complexity. Accordingly, a language is increasingly isolating as it has less and less morphological structure, increasingly monocategorial as its syntactic categories decrease in number and importance, and increasingly associational as its construction-specific rules of semantic interpretation become fewer and less distinct. Alongside **Pure IMA Language**, as described in (1) above, one may thus entertain the possibility of a range of **Relative IMA Languages**, approaching Pure IMA Language to various degrees within each of the three domains.
2.1. Isolating

The first defining property, *morphologically isolating*, is the one that is most familiar, since it pertains to a typology that has been the focus of considerable attention in the linguistic literature. As is well known, isolating languages such as Vietnamese have considerably less word-internal morphological structure than synthetic languages such as Russian, which in turn have considerably less morphology than polysynthetic languages such as Mohawk. However, no natural language is purely isolating, as per (1a); all known isolating languages still have *some* morphology – affixation, compounding, or other kinds of processes such as reduplication, stem alternation, and so forth.

In a purely isolating language, without any morphology whatsoever, there would be no distinction between words and morphemes: every word would contain exactly one morpheme, and every morpheme would constitute a word. There would thus be no need to maintain both concepts; one of the two could be discarded. In fact, Chomsky’s (1965) *Aspects of the Theory of Syntax* makes a similar proposal for English; in this model, the terminal nodes of syntactic trees, called “formatives,” are actually morphemes, and the notion of word is done away with entirely. However, as shown by Anderson (1982), such a model is empirically inadequate, since, in English and presumably all natural languages, the ways in which morphemes are put together to form words are fundamentally different from the ways in which words are grouped together to form sentences. However, it is worth keeping in mind that although there are no purely isolating languages, this is a fact about the way languages are, not about how they must necessarily be: it is not difficult to imagine a hypothetical language in which the smallest meaning-bearing units all behaved as independent words, grouping together according to syntactic principles.

2.2. Monocategorial

The second defining property, *syntactically monocategorial*, pertains to a domain within which the presence of crosslinguistic variation has only recently, and still only partially, been recognized. In the past, syntactic categories have generally been presumed to be universal, often in accordance with the eight parts of speech of traditional Latin grammar. Indeed, the assumption that syntactic categories must be the same in all languages has lingered on into much current linguistic work, in schools as diverse as linguistic typology and generative grammar; this assumption is in evidence whenever a linguist analyzing a language says that one word must be a noun because it means ‘chicken’ while another word must be a verb because it means ‘eat.’

However, in recent years an increasing body of literature has begun to examine the ways in which the inventories of syntactic categories may vary across languages [see, for example, Gil (2000b)]. One important issue that has attracted considerable attention has been the viability and nature of the category of adjective, the extent to which words denoting properties such as ‘big,’ ‘red,’ ‘good’ and so forth exhibit distinct adjectival behavior, or, alternatively, are subsumed within larger categories of noun or verb [see, for example, Dixon (1977), Stassen (2005), Wetzer (1992), and others]. Another major focus has been
on the universality of what is generally considered to be the most fundamental categorial
distinction, namely that between noun and verb; such work has typically dealt with lan-
guages which seem, prima facie, to lack a noun/verb distinction, from families such as
Wakashan [Swadesh (1939)]. It is, of course, languages lacking a noun/verb distinction
which come closest to being syntactically monocategorial. However, to the best of my
knowledge, no language has ever actually been proposed to be purely monocategorial. In
particular, most or all descriptions of languages without a noun/verb distinction still
involve, at the very least, a distinction between a single open syntactic category (encom-
passing the equivalents of both nouns and verbs) and one or more closed syntactic cate-
gories containing various “grammatical” or “functional” items.

2.3. Associational

The third defining property, semantically associational, although rooted in various
commonplace observations concerning the ways in which expressions derive their
meanings, is nevertheless of a more novel nature. Consider the best translation of a
basic transitive sentence such as Mary hit John into the language of your choice. How
do you know who hit whom? If you chose Mandarin, then, as in English, the agent is
differentiated from the patient by linear order: the agent precedes the verb while the
patient follows it. However, if you chose Russian, then linear order provides no seman-
tic information; instead, the agent is differentiated from the patient by its case marking,
nominative as opposed to accusative, and by the fact that it triggers gender agreement
on the past-tense form of the verb. The various rules whereby agents and patients are
differentiated in English, Mandarin, Russian, and other languages constitute examples
of construction-specific rules of semantic interpretation, as specified in (1c) above, in
that they apply specifically to active transitive clauses. Most languages contain many
such construction-specific rules, which, together, govern the compositional semantics
of clauses, phrases, and other, more specific constructions, accounting for semantic fea-
tures such as thematic roles, tense, aspect, number, definiteness, and numerous others.

Now imagine that, armed only with a rudimentary dictionary, you are confronted
with a three-word sentence in an unfamiliar language. Somehow, you identify three
word stems, meaning ‘Mary,’ ‘hit,’ and ‘John’; however, these three word stems bear
rich additional morphological structure, and you know nothing about the grammar of
the language. Can you figure out the meaning of the sentence? At first blush, the answer
would seem to be no. With no information on thematic roles, tense, aspect, number, def-
initeness, and other such features, the sentence could mean anything from ‘Mary hit
John’ through ‘John will repeatedly try to hit Mary’ to ‘John and Mary aren’t hitting
anybody’ and so on and so forth. Still, the meaning of the sentence is hardly uncon-
strained: it is not very likely to mean ‘The rain in Spain falls mainly in the plains.’ Thus,
although you have no knowledge of the grammar of the language, it is a safe bet, in fact
a near certainty, that the meaning of the sentence, whatever it is, has to do in some way
with ‘Mary,’ ‘hit,’ and ‘John.’
The semantic relationship of “having to do with” may be formally represented by means of the Association Operator, defined as follows:

(2) The Association Operator $A$:
Given a set of $n$ meanings $M_1 \ldots M_n$, the Association Operator $A$ derives a meaning $A(M_1 \ldots M_n)$ read as ‘entity associated with $M_1$ and $M_n$.’

Two subtypes of the Association Operator may be distinguished: the Monadic Association Operator, in which $n = 1$, and the Polyadic Association Operator, for $n > 1$.

In its monadic variant, the Association Operator is familiar from a wide variety of constructions in probably all languages. Without overt morphosyntactic expression, it is manifest in cases of metonymy such as the often cited *The chicken left without paying*, where the unfortunate waiter uses the expression *the chicken* to denote the person who ordered the chicken. Using small uppercase letters to represent the meanings of individual expressions, we can represent the meaning of *chicken* in the above sentence by means of the Monadic Association Operator as $A(\text{CHICKEN})$, or ‘entity associated with chicken.’ The nature of the association between the entity and the chicken is left open by the Association Operator, to be filled in by the context, which, in the case of a restaurant, is the obvious one involving a dishonest or forgetful customer. Similar examples are everywhere. In the *International Herald Tribune*, on 5–6 November, 1994, a newspaper headline reads *Washington Turns Away From Japan Trade Fight*, with a subhead *Clinton Planning to Shift the Emphasis To Markets in Asia and Latin America*. The continuation of the article makes it clear that the expressions *Washington* and *Clinton* are to be understood metonymically, as $A(\text{WASHINGTON})$, ‘entity associated with Washington,’ and $A(\text{CLINTON})$, ‘entity associated with Clinton,’ which, in the context of the article, both denote the Clinton administration.

Even more often than the above cases, the Monadic Association Operator is overtly expressed via a specific form, which is commonly referred to as a genitive, possessive, or associative marker. Consider, for example, the English possessive enclitic ‘s. Application of ‘s to *John* yields the expression *John’s*, which has the interpretation $A(\text{JOHN})$, ‘entity associated with John,’ where the nature of the association is unspecified. Some idea of how unconstrained the association is can be obtained by comparing the obvious meanings of phrases such as *John’s father*, *John’s nose*, *John’s shirt*, *John’s birthday*, *John’s suggestion*, and so forth, or by considering the range of meanings of a single phrase such as *John’s book*, which could denote the book that John owns, the book that John wrote, the book that is about John, or, in more specific contexts, the book that John was assigned to write a review of, and so forth. Another example of an association marker is provided by the Mandarin form *de*. In Mandarin, the expression *Yuèhàn de* has more or less the same range of interpretations as the English *John’s*. However, unlike English ‘s, Mandarin *de* can apply to expressions belonging to other syntactic categories, as for example in *Yuèhàn mài de*, where *Yuèhàn mài* means ‘John buy.’ Many descriptions of Mandarin characterize constructions such as these as relative clauses, and assign them translations such as ‘One that John bought.’ However, the same expression can also mean ‘the manner of John’s buying,’ ‘the extent of John’s buying,’ and so forth. Moreover, characterizing *de* as a genitive marker in *Yuèhàn de* but at the same time as a
relative clause marker in Yēhàn māi de misses an obvious generalization. Specifically, in both cases, de shares the function of an associative marker. While in the former case, Yēhàn de has the interpretation A (JOHN), ‘entity associated with John,’ in the latter case, Yēhàn māi de is assigned the interpretation A (JOHN.BUY), ‘entity associated with John’s buying,’ an interpretation which underlies all of the available readings of the expression in question. As suggested by the two examples considered above, association markers may differ from each other in their syntactic properties. Indeed, even within English, association markers may differ from each other syntactically and also semantically, as evidenced by the numerous and sometimes subtle contrasts between the enclitic ’s and the other associative marker of. Such contrasts suggest that the denotations of such markers, although based on the Association Operator, may involve additional and sometimes more idiosyncratic semantic components.

In its polyadic variant, the Association Operator provides for a basic mechanism of compositional semantics in which the meaning of a complex expression is derived from the meanings of its constituent parts. In accordance with the Polyadic Association Operator, whenever two or more expressions group together to form a larger expression, the meaning of the combined expression is associated with, or has to do with, the meanings of each of the individual expressions. Obviously, polyadic association applies in a default manner throughout language; it is hard to imagine how things could be otherwise. Thus, in the little thought experiment described above, it is what made it possible to be sure that in an unfamiliar language, in the absence of any specific grammatical information, in a sentence with three words whose meanings were based on ‘Mary,’ ‘hit,’ and ‘John,’ the meaning of the sentence would still be associated in some way with ‘Mary,’ ‘hit,’ and ‘John,’ or, in terms of the Polyadic Association Operator, A (MARY, HIT, JOHN), ‘entity associated with Mary, hitting, and John.’

One grammatical domain in which the Polyadic Association Operator is overtly visible is in genitive constructions. In many languages, genitive constructions are formed by the bare juxtaposition of the two expressions, in which case the derived meaning may be represented by means of the Polyadic Association Operator applying without any overt morphosyntactic expression. For example, in Yagua, a language isolate of northeastern Peru, Tomáása rooriy has a range of interpretations resembling that of its English translation ‘Tom’s house’ [Payne and Payne (1990), p. 348]; its meaning may thus be represented as A (TOM, HOUSE), ‘entity associated with Tom and house.’ A similar mechanism generates the set of potentially available meanings of nominal compounds in English and other languages, though their actual meanings are usually the product of further arbitrary conventionalization. Thus, for example, a magnifying glass is an instrument of magnification, a looking glass is a goal of looking, a sherry glass is a receptacle of sherry, and so on. However, even in the above examples, the actual meanings are somewhat more specific than those derived by the Polyadic Association Operator; they therefore fail to be purely associational. Thus, in Yagua, genitive constructions are right-headed; hence, the ‘entity associated with Tom and house’ has the further property of being, itself, a house; a more accurate translation would have been ‘house associated with Tom.’ Similarly, English compounds such as those listed above are also right-headed; accordingly, in these cases too, the first element is construed as
modifying the second one. (Some of the ways in which semantic representations based on the Polyadic Association Operator are narrowed down by the imposition of headedness are discussed in more detail in Section 4.2.3 below.)

More generally, the Polyadic Association Operator may be considered as a universal default mechanism for semantic interpretation, but one that is in most cases overridden and narrowed down substantially by the application of additional construction-specific rules. A purely associational language would be one in which there were no such further construction-specific rules of semantic interpretation, and in which, therefore, the compositional semantics was effected exclusively by the Polyadic Association Operator. It is almost certainly the case that no natural language is purely associational; however, as argued in Section 4 below, some languages may come closer to being purely associational than is generally assumed.

In general, Pure IMA Language represents a limiting case of maximal simplicity within the domains of morphology, syntax, and semantics. One may indeed wonder whether IMA Language is capable of fulfilling the multifarious functions associated with human language in the diverse contexts in which it is used. Nevertheless, as we shall now see, IMA Language is in fact more widespread than might be expected, and can indeed fulfill a wider range of functions than might seem, prima facie, to be the case.

3. Where IMA Language Is Found

IMA Language, or a system that comes close to it, is manifest in the following five distinct ontological realms:

(3) (a) Semiotics
   Some artificial languages are IMA Language;

(b) Phylogeny
   At some stage in evolution, early language was IMA Language;

(c) Ontogeny
   At some stage in acquisition, early child language is IMA Language;

(d) Typology
   Some languages come closer than others to IMA Language;

(e) Cognition
   IMA Language is a feature of general human cognition.

The first three domains – semiotics, phylogeny and ontogeny – are considered briefly in this section; the fourth, typology, is discussed in greater detail in Section 4; while the fifth, cognition, is taken up in Section 5.

3.1. Semiotics

IMA Language may be observed in a variety of artificial semiotic systems. One such system is the language or languages of pictograms, those familiar iconic signs that can
be seen in airports, railway stations, and many other places, including in particular those that have the specific function of traffic signs. To see how pictograms instantiate IMA Language, let us consider a typical example of pictogram usage: the juxtaposition of two signs, one consisting of an arrow, the other depicting a bicycle, as represented below:

![Pictogram example](image)

Clearly, the language of pictograms is compositional, since we can take simple signs and combine them to form more complex signs. Nevertheless, there would seem to be no evidence for any distinction between different compositional systems corresponding to morphology and syntax in natural languages. Under the most obvious analogy, the arrow and the bicycle picture are the equivalents of words, while the combination of the two signs belongs to syntax; however, neither of the two signs has any internal meaning-bearing structure of the type that might then be characterized as morphological. (Whatever internal structure the bicycle sign may exhibit does not qualify, since such structure is inherent to the iconic nature of the sign; it would not make sense to characterize, say, the line depicting the handlebar as an individual morpheme, since it occurs in no sign other than the bicycle icon.) Accordingly, in the absence of anything corresponding to inflection or polysynthesis, the language of pictograms may be considered to be morphologically isolating.

Similarly, the language of pictograms would appear to be devoid of any evidence for distinct syntactic categories. In the above example, the arrow and bicycle signs belong to the same “part of speech” – in fact the only one in the language of pictograms. More generally, there are no noun signs, adjective signs, verb signs, or any other syntactic categories of signs. All signs have the same distributional privileges: any two or more signs may be juxtaposed without any constraints of the kind that are reflected in the familiar grammaticality judgments of ordinary natural languages. Thus, the language of pictograms may also be viewed as monocategorial.

But what about the meaning of our pictogram example of the arrow and bicycle? In many European cities, the most common meaning of such a collocation, one that has undergone a certain degree of conventionalization, is to denote a special bicycle lane: ‘bicycles go thataway.’ However, in at least one case, I have observed a similar combination used to point the way to a bicycle shop: ‘go thataway for bicycles.’ Thus, the example would seem to be vague or ambiguous. In terms of the categories of natural languages, the arrow would seem to denote an activity which may assign a thematic role to the bicycle: agent in the former meaning, goal in the latter. However, there is no reason internal to the language of pictograms to posit the existence of thematic role assignment of any kind. Rather, by means of the Polyadic Association Operator, we may represent a general unified meaning underlying the two more specific ones, with the formula A (BICYCLE, THATAWAY), ‘entity associated with bicycle and with thataway,’ where the details of the association are filled in by the context. In general, whenever we encounter two signs in
close proximity, we assign the combination a meaning that has to do in some way with
the meanings of the individual signs, in accordance with the Polyadic Association
Operator. Accordingly, the language of pictograms may also be viewed as associational.

In sum, then, the language of pictograms satisfies the three properties of IMA
Language. Clearly, the language of pictograms does not have the entire range of expres-
sive power associated with ordinary natural languages. Nevertheless, one can still say
quite a lot with pictograms, and their functionality is boosted by a substantial reliance
on context: whether our example pictogram is intended to mean ‘bicycles go thataway,’
‘go thataway for bicycles,’ or perhaps something else again, can readily be inferred by
the location of the sign, supported by various other contextual cues.

3.2. Phylogeny

Although we have precious little direct evidence of any kind concerning the evolu-
tion of natural language, it is reasonable to suppose that early human language was
IMA Language. More precisely, the following two logically distinct hypotheses may be
formulated:

(4) (a) *Evolution of Linguistic Abilities*

At some stage in evolution, the cognitive abilities of humans or prehumans
were limited to the representation of IMA Language;

(b) *Evolution of Actual Languages*

At some stage in evolution, all natural languages were IMA Language.

While hypothesis (4a) is about the evolution of cognition, or, more specifically, men-
tal grammar, sometimes referred to as I-language, hypothesis (4b) is about the evolu-
tion of actual languages, also known as E-languages.

A commonly held position, most often associated with Chomsky and his followers, is
that contemporary human linguistic abilities emerged *ex nihilo* in a single gigantic leap, pre-
sumably associated with a unique genetic mutation. Such a view is clearly inconsistent with
hypothesis (4a); however, it is agnostic with respect to hypothesis (4b), since even if human
linguistic abilities went straight from nothing to what they are now, actual languages might
have taken a variety of incremental paths over the course of time in order to make use of
such abilities (indeed this process may still be far from complete); and one of those possi-
bile paths could easily have involved IMA Language as an evolutionary way station.

A more refined position is put forward by Bickerton (1990), who argues that human
linguistic abilities evolved into their contemporary shape through an intermediate stage
which he refers to as *protolanguage*. Structurally, Bickerton’s protolanguage is a form
of IMA Language; however, it embodies at least one significant further restriction that
is not part of IMA Language, namely that it does not permit syntactic recursion.
Ontologically, too, Bickerton’s protolanguage is akin to IMA Language, in that he con-
siders it to be manifest in a variety of realms, including three of the five listed in (3)
above: phylogeny, ontogeny, and cognition. Notably, however, Bickerton has nothing to
say about the other two domains: semiotics and typology. Moreover, he expressly
denies the existence of any “interlanguage” between protolanguage and contemporary
linguistic abilities; thus, like Chomsky, his position is inconsistent with hypothesis (4a),
though in the case at hand, what is at issue is a single, albeit very important structural
feature, namely, syntactic recursion. Conversely, hypothesis (4a) is consistent with, but
does not necessarily entail, the existence of a stage, prior to IMA Language and the evolu-
tion of recursion, corresponding to Bickerton’s protolanguage.

So how might we seek support for the two evolutionary hypotheses in (4)? Although
we cannot go back in time, we can jump across the branches of our evolutionary tree to
see what our nearest relatives, the various primates, have accomplished in the realm of
language. Many species have a lexicon of predator cries; however, since these usually
involve individual cries in isolation, there is no compositionality, and hence nothing near
the possible richness of IMA Language. A somewhat more interesting case, reported
recently by Zuberbuhler (2002), is that of male Campbell’s monkeys, who appear to be
able to juxtapose two different calls, a predator cry preceded by a “boom” sound, to pro-
duce a complex cry whose meaning seems to involve some kind of attenuation or even
negation of the predator-cry meaning. However, to this point at least, no clear examples
of productive compositionality of meaning-bearing signs have been attested in the natu-
really occurring repertoire of nonhuman primates, or any other animals.

However, among primates in captivity, there is an increasing body of evidence sug-
gesting that they can be taught to master compositionality, and concomitantly also IMA
Language. Two of the more celebrated cases are those of the bonobo Kanzi [Greenfield
and Savage-Rumbaugh (1990)], using lexigrams, and the orangutan Chantek [Miles
(1990)], using American sign language. Some examples of Kanzi’s spontaneous lexi-
gram production are given below:

(5) (a) LIZ HIDE  agent – HIDE
(b) WATER HIDE  patient – HIDE
(c) HIDE AUSTIN  HIDE – agent
(d) HIDE PEANUT  HIDE – patient

Kanzi’s usage of lexigrams provides no evidence for morphological structure or for dis-
tinct syntactic categories; thus, it is isolating and monocategorial. Moreover, as suggested
by examples such as those in (5), it is also associational. The above examples form a minia-
ture paradigm (schematized to the right) in which the same sign HIDE is either preceded or
followed by a participant, which, as indicated by the context of the utterance given by the
authors, may, in either position, be understood as either the agent or the patient. Thus, there
would seem to be no evidence for any grammatical assignment of thematic roles in Kanzi’s
lexigram usage. Rather, the semantic relationship between the two signs is vague. As in the
language of pictograms and the arrow-and-bicycle example, the juxtaposition of lexigrams
has a single general meaning that may be represented in terms of the Polyadic Association
Operator as, for (5a), A (LIZ, HIDE), ‘entity associated with Liz and with hiding.’ Thus, the
bonobo Kanzi’s use of lexigrams satisfies the three properties of IMA Language. Similar
observations hold also for the orangutan Chantek’s usage of ASL.
It would seem, then, to be the case that both bonobos and orangutans are endowed with the cognitive abilities to represent IMA Language, even though they apparently have not made any use of these abilities to create any actual IMA Languages in the wild. Given that the common evolutionary ancestor of bonobos and orangutans is also an ancestor of humans, it is likely that this common ancestor also had the cognitive abilities to represent IMA Language without having any actual IMA Languages. (The alternative, less parsimonious scenario would involve positing the independent development of IMA Language abilities in at least two separate evolutionary lineages.) Quite obviously, however, no primates, even in captivity and with the dedicated efforts of their caregivers, are capable of acquiring the full-blown complexities of natural human language. Thus, the linguistic capabilities of captive apes support the reconstruction of a stage in human evolution, perhaps eight or ten million years ago, in which the ability to represent IMA Language was already present, in accordance with hypothesis (4a). The linguistic capabilities of captive apes also increase the plausibility of hypothesis (4b), though the alternative logical possibility remains that prehuman cognitive abilities may have developed past IMA Language before actual languages ever reached the IMA stage.

It should be noted, though, that since, to the best of my knowledge, the linguistic behavior of captive apes does not provide any evidence for the mastery of syntactic recursion, the abilities of Kanzi, Chantek, and other such captive apes may equally well be characterized in terms of Bickerton’s more restrictive protolanguage. In order to provide specific support for the existence of an evolutionary stage of IMA Language, either in addition to or instead of protolanguage, evidence of a different kind is called for: at present I am not familiar with any such evidence.

3.3. Ontogeny

As ontogeny is said to recapitulate phylogeny, IMA Language may also be observed in early child language. Again, whereas a nativist position, associated with Chomsky and his supporters, holds that all the complexity of adult language is, in some form or guise, present from the outset, alternative approaches to first-language acquisition point toward the more commonsensical position that language does indeed develop as the child grows older. And indeed, it would seem to be the case that children pass through a stage in which they have acquired a system resembling that of IMA Language.

In the domain of morphology, there is ample evidence that children acquiring a language with rich morphology start out by treating individual words as unanalyzable wholes, only later becoming aware of their internal structure. Thus, early child language may be characterized as isolating.

In syntax, it would seem to be the case that early child language lacks distinct syntactic categories. If such categories are defined distributionally, then of course at the one-word stage, early child language is monocategorial by definition, since all words occur in the same one-word construction. However, there is reason to believe that monocategoriality may extend also into the two- or multiword stage. In Gil (2000b), a categorial-grammar-based theory of syntactic categories is proposed which, among other things, suggests
specific hypotheses about the order in which syntactic categories are acquired. And in Gil (2003), empirical support for this theory is provided from a study of the acquisition of Jakarta Indonesian, in which, it is argued, children pass through a monocategorial but multiword stage before acquiring an additional distinct syntactic category.

Semantically, it has also been suggested that early child language is lacking in many or all of the construction-specific rules of semantic interpretation characteristic of adult language. Consider the following two examples, cited by Bloom (1973), from the spontaneous speech of Allison, at age 20 months, who is playing with a pig inside a toy truck; the pig is hurt by a sharp corner of the truck:

(6) (a) hurt truck  \textit{HURT} – cause
(b) hurt knee  \textit{HURT} – patient

Like Kanzi’s lexigram examples in (5), the above utterances form a mini-paradigm (indicated to the right) in which \textit{hurt} is followed by a participant, which, as suggested by the context of the utterance, may be understood as either the cause or the patient. Accordingly, Bloom argues that there is no justification for reading into utterances such as these any kind of grammatical structure involving relations such as subject and object which determine thematic roles. Rather, the semantic relationship between the two words is underspecified. As in the language of pictograms and the signs of captive apes, the juxtaposition of words in early child language may thus be attributed a single general meaning represented in terms of the Polyadic Association Operator as, for (6a), \textit{A (HURT, TRUCK)}, ‘entity associated with hurt and with truck.’ Accordingly, early child language may also be characterized as associational.

Thus, early child language passes through a stage in which it exhibits the properties of IMA Language, before moving on to develop further complexity. Note that at the two-word IMA stage, early child language is syntactically nonrecursive; at this stage, then, the child’s linguistic abilities instantiate the more restricted system of protolanguage, as indeed Bickerton (1990) points out. Whether the child retains IMA Language after acquiring syntactic recursion is an issue that needs to be further explored; in fact, it is possible that the answer to this question may vary with the choice of target language, in accordance with the language’s own typological profile.

4. Typology

As suggested above, the structural properties of IMA Language are shared by three quite distinct ontological realms: artificial semiotic systems, early evolutionary stages of language, and the language of young children. However, as noted earlier, most natural languages exhibit a much greater degree of complexity than is characteristic of IMA Language.

From a typological perspective, natural languages may vary independently in the extent to which they exhibit each of the three defining features of IMA Language. Of these three features, however, only the first, pertaining to morphological structure, is readily observable in a relatively theory-neutral way; the remaining two, pertaining to syntactic and
semantic structure, presuppose in-depth linguistic analysis, which may vary in its conclusions in accordance with the theoretical persuasions of the linguist conducting the analysis. Accordingly, in the present state of the art, we cannot really compare languages, but only descriptions of languages, each as seen through the eyes of a different linguist armed with different theoretical assumptions and using different research methodologies. Further compounding the problem, most linguistic descriptions, of whatever orientation, exhibit a bias in favor of positing more syntactic categories and construction-specific semantic rules than are actually warranted by the facts of the language in question [see Gil (2001b) for a discussion of the pervasive Eurocentrism that underlies this bias]. Accordingly, existing descriptions of languages tend toward a systematic underestimation of the degree to which the properties of IMA Language are approximated by individual languages.

With these qualifications in mind, one may nevertheless engage in some elementary comparisons of the IMA properties of different languages. Russian, under any standard description, is as far from IMA Language as one can get: it has rich morphological structure, well-motivated syntactic categories, and lots of construction-specific rules of semantic interpretation. Vietnamese, in accordance with most descriptions [for example, Thompson (1965)] is strongly isolating; however, it is characterized as possessing distinct syntactic categories and construction-specific rules of semantic interpretation. Tagalog, as argued by Gil (1993), comes close to being monocatational, with but a single open syntactic category; however, it clearly has rich morphological structure and a variety of construction-specific semantic rules. Given the logical independence of the three IMA properties, collapsing their scalar nature into an idealized binary “high/low” distinction would yield a total of eight different language types. However, at present, whether all eight language types are actually attested must remain unanswered.

Nevertheless, as we shall now see, some natural languages may come surprisingly close to exhibiting the three properties characteristic of IMA Language. The following subsections present a more detailed exploration of one particular exemplar of a Relative IMA Language: the Riau dialect of Indonesian.

4.1. Riau Indonesian: overview

Riau Indonesian is the variety of Malay/Indonesian spoken in informal situations by the inhabitants of Riau province in east-central Sumatra, Indonesia; it is quite different from standard Indonesian, which is familiar to many general linguists from a substantial descriptive and theoretical literature. Riau Indonesian is one of a number of regional varieties of colloquial Indonesian, which, although different from each other in numerous details, nevertheless share the same typological ground plans. (One such regional variety is Jakarta Indonesian, referred to briefly in Section 3.3 above.) Thus, the characterization of Riau Indonesian as a Relative IMA Language is probably applicable to a wide range of colloquial varieties of Indonesian, totaling tens of millions of native speakers.

The first IMA property, morphologically isolating, clearly applies to a very great extent to Riau Indonesian [see Gil (2002a, 2004a, to appear) for description and analysis of various aspects of Riau Indonesian word structure]. Inspection of any text will
reveal a low word-to-morpheme ratio, as well as substantial stretches in which the word-to-morpheme ratio is actually one-to-one. Riau Indonesian has no inflectional morphology whatsoever, and little in the way of derivational morphology. Only three productively used items are clearly affixal: the prefixes se- ‘one’; (s)i-, marking names of persons; and N-, marking agent-orientation. (Actually, the latter form is prefixal in only some of its allomorphs; it is otherwise sometimes realized as a proclitic me-) In addition, there are a handful of items, probably fewer than ten, whose nature is intermediate between affixes and clitics; among these are the forms ber-, marking nonpatient orientation; ke-, marking direction; and –an, which has a variety of usages that may or may not be related to each other. Some other items whose cognates are attached to their hosts in written Standard Indonesian are clearly clitics rather than affixes in Riau Indonesian; these include the forms di-, marking patient orientation; ter-, marking nonagent orientation; and –kan, marking end-point orientation.

In fact, the most commonly occurring bound morphemes in Riau Indonesian are actually suprasegmental rather than linear. The most important of these is reduplication – usually complete though sometimes partial – which has a variety of usages [see Gil (2005a) for detailed discussion and analysis]. Another is truncation, used productively to create familiar forms from names, e.g., Ril from Kairil, or from other terms of address, e.g., bang from abang ‘elder brother.’ Finally, like most or all languages, Riau Indonesian makes use of compounding, though with two important qualifications. First, the two terms of the compound are less strongly bound to each other than in many other languages; in fact, there would appear to be no phonological grounds for distinguishing between compounds and phrasal collocations (corresponding to, say, the stress shift that is evident in English, or the construct-state inflection that is characteristic of Hebrew). Secondly, compounds appear to be less common than in many other languages. This is particularly striking in comparison with the isolating languages of mainland Southeast Asia: Chinese, Thai, Vietnamese, and so forth. Although the latter languages are traditionally thought of as monosyllabic, a substantial body of recent phonological literature suggests that they are also characterized by a bisyllabic minimal word [see, for example, Bao (1990), Yip (1991), and Feng (2002) on Sinitic languages]. And indeed, one of the most productive devices for achieving the canonical bisyllabic word in such languages is compounding. However, in Riau Indonesian, the canonical monomorphic word is already bisyllabic, and possibly for this reason, compounding occurs much less frequently. Indeed, the scarcity of compounding in Riau Indonesian in contrast to the monosyllabic languages of mainland Southeast Asia suggests that Riau Indonesian may actually represent an even more extreme case of an isolating language.

But what of the other two IMA properties? In order to evaluate the extent to which these apply, we need to take a deeper look into the syntactic and semantic patterns of Riau Indonesian. [Much of what follows in Sections 4.1 and 4.2 below is an abridged version of a more detailed discussion and analysis presented in Gil (2005b)]. As a point of departure, we shall take the following English sentence.

(7) The chicken is eating.
How might one go about translating the above sentence into Riau Indonesian? Two natural and idiomatic translations are given in (8):

(8) (a) Makan ayam  
        eat chicken

(b) Ayam makan  
        chicken eat

‘The chicken is eating’

Sentences (8a) and (8b) each consist of two monomorphemic words, ‘eat’ and ‘chicken’: the only difference between them involves word order. They mean the same thing, and they are equally natural, though their pragmatic appropriateness conditions differ somewhat.

How similar are (7) and (8)? A pedagogical grammar of Malay (quoting an anonymous source) has the following to say:

...the Malay and English sentence structures are so similar that one scholar has even remarked that “Indonesian (in this case Malay), is a western language using Indonesian (Malay) words.”

Liaw (2002), p. iv

This view is implicit in typologies such as that of Greenberg (1963), who classifies the world’s languages into 24 different word-order types, and puts Malay into the same cell as many European languages, including the Romance languages and Modern Greek. And it is explicit in much recent work on Malay/Indonesian within the generative framework, such as Guilfoyle, Hung and Travis (1992), which proposes syntactic structures that are well-nigh indistinguishable from those of English.

It takes a novelist, albeit one with great linguistic sensitivity, to see the obvious differences between Malay and English, which have escaped the sight of so many linguists. Here is the perspective of Anthony Burgess:

What strikes the learner of Malay is the complete lack of those typically Indo-European properties – gender, inflection, conjugation. It is like diving into a bath of pure logic. Everything is pared to a minimum. [...] If one digs deeply enough into Malay, one comes to the conclusion that the Western concept of “parts of speech” is alien to it.

[Burgess (1975), pp. 183–184]

Anthony Burgess was right on. Let us now take Burgess up on his suggestion and begin digging.

Table 1 summarizes some of the more salient differences between sentence (7) in English and its two translations into Riau Indonesian in (8).

The first difference in Table 1 is a formal one. English sentence (7) exhibits numerous structural asymmetries. Two morphosyntactic asymmetries are easily visible on the surface: the NP the chicken controls agreement of the auxiliary is, and the auxiliary is in turn governs the -ing ending on the verb. Lurking beneath these morphosyntactic asymmetries are a host of syntactic asymmetries, providing the motivation for grammatical analyses of sentences such as (7) as involving subject and predicate, NP and VP, or whatever. In contrast, the Riau Indonesian sentences in (8) are completely symmetric, their two
constituent parts being totally balanced. There is no morphological agreement or government, either in (8) or anywhere else in Riau Indonesian. Moreover, the lack of morphological asymmetry mirrors the absence of any deeper syntactic asymmetry. As argued elsewhere, words such as *makan* ‘eat’ and *ayam* ‘chicken’ have the same distributional privileges and, more generally, identical syntactic behavior. They therefore belong to the same syntactic category, in fact the only open syntactic category in Riau Indonesian, namely S. Structurally, then, the two Riau Indonesian sentences in (8) are instances of sentential coordination, with a structure of the form [S S S], as represented in (10) below.

Table 1

<table>
<thead>
<tr>
<th>Symmetry</th>
<th>English</th>
<th>Riau Indonesian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>asymmetric:</td>
<td>symmetric</td>
</tr>
<tr>
<td>Agreement (on CHICKEN)</td>
<td>The chicken → is</td>
<td></td>
</tr>
<tr>
<td>Government (on is)</td>
<td>is → -ing</td>
<td></td>
</tr>
</tbody>
</table>
| Number (on CHICKEN)        | marked: singular | unmarked: also means
|                          | ‘The chickens are eating’ |
| Definiteness (on CHICKEN)  | marked: definite | unmarked: also means
|                          | ‘A chicken is eating’ |
| Tense (on EAT)             | marked: present | unmarked: also means
|                          | ‘The chicken was eating’
|                          | ‘The chicken will be eating’ |
| Aspect (on EAT)            | marked: progressive | unmarked: also means
|                          | ‘The chicken eats’
|                          | ‘The chicken has eaten’ |
| Thematic role (on CHICKEN) | marked: agent | unmarked: also means
|                          | ‘Someone is eating the chicken’
|                          | ‘Someone is eating for the chicken’
|                          | ‘Someone is eating with the chicken’ |
| Ontological type (on CHICKEN EAT) | marked: activity | unmarked: also means
|                          | ‘The chicken that is eating’
|                          | ‘Where the chicken is eating’
|                          | ‘When the chicken is eating’ |

The remaining differences presented in Table 1 are semantic. In the English sentence in (7), the subject NP is marked for number and definiteness, like most other NPs in English. In contrast, in the Riau Indonesian sentences in (8), *ayam* ‘chicken’ is unmarked for number and definiteness; number marking is almost completely absent in Riau Indonesian, while definiteness marking is optional. Thus, Riau Indonesian (8) has a wider range of interpretations than its English counterpart, as suggested by the additional translations of (8) back into English in Table 1. Similarly, in English (7), the verbal phrase *is eating* is marked for tense and aspect, like most other verbal phrases in English. In contrast, in Riau Indonesian (8), *makan* ‘eat’ is unmarked for tense and aspect; these two categories are expressed in this language by optional periphrastic devices which are
for the most part only weakly grammaticalized. Once more, Riau Indonesian (8) has a wider range of interpretations than its English counterpart, as suggested by the many ways in which (8) can be translated into English.

Whereas the absence of number, definiteness, tense, or aspect marking is a familiar areal feature of Southeast Asian languages, the remaining two characteristics of Riau Indonesian are perhaps somewhat more exceptional from a crosslinguistic point of view. In English (7), the NP the chicken is marked as bearing the thematic role of agent. In general, thematic roles are central to the grammatical organization of English and of many other languages. In contrast, in Riau Indonesian (8), the expression ayam ‘chicken’ is not marked for thematic role; as suggested by the alternative translations of (8) back into English in Table 1, ayam ‘chicken’ could also be interpreted as patient, or, given an appropriate context, as benefactive, comitative, or any other thematic role whatsoever. The indeterminacy of thematic roles in Riau Indonesian is exemplified and discussed in detail in Gil (1994, 2002b), and is argued in Gil (2001b, 2005b) to be an instance of vagueness rather than ambiguity.

The final difference between (7) and (8) presented in Table 1 is perhaps the most fundamental one; it pertains to the ontological type of the expressions. Whereas English (7) denotes an activity, Riau Indonesian (8) is unmarked for ontological type. Again, Riau Indonesian (8) has a wider range of interpretations than its English counterpart, as evidenced by the additional translations of (8) back into English. As suggested by these translations, the sentences in (8) could also denote a thing (‘The chicken that is eating’), a place (‘Where the chicken is eating’), a time (‘When the chicken is eating’), and so on. Again, the indeterminacy of ontological types in Riau Indonesian is exemplified and discussed in detail in Gil (2001b, 2005b), where it is also argued that this indeterminacy is an instance of vagueness rather than ambiguity.

Thus, as summarized in Table 1, the Riau Indonesian sentences in (8) differ fundamentally from their English counterpart in English (7); in fact, they bear a much greater resemblance to the arrow-and-bicycle pictogram, Kanzi’s lexigram usage in (5), and the early English child-language utterances in (6). Formally, the Riau Indonesian sentences lack any evidence of asymmetrical structure; semantically, they are unmarked and in fact vague with respect to the categories of number, definiteness, tense, aspect, thematic role, and ontological type. Anthony Burgess was right; this is, indeed, a language “pared to a minimum.” In fact, this minimum is one that provides a relatively close approximation to IMA Language.

4.2. Riau Indonesian: analysis

We shall now sketch the outlines of Riau Indonesian syntax and semantics, proposing explicit representations for the observations made above. In doing so, we shall continue to use the two sentences in (8) as a convenient point of reference.

4.2.1. Syntax

As argued in Gil (1994, 2000b, 2001b), in Riau Indonesian there is but a single open syntactic category: S, or sentence. All members of S exhibit the same syntactic behavior,
including the same distributional privileges. In particular, all members of S can stand alone as complete nonelliptical sentences. The category S includes makan ‘eat’ in (8), and practically all other words whose translational equivalents in English are verbs, and also ayam ‘chicken’ in (8), and just about all other words whose translational equivalents in English are nouns. In addition, the category S includes most words whose translational equivalents in English are adjectives, prepositions, and determiners, plus a variety of words whose closest English counterparts are function words or morphemes. Such words include tak, marking negation; udah, denoting the perfect; sendiri, which expresses a variety of notions including restrictive focus, intensification, and reflexivity [Gil (2001c)]; and sama, whose usages range over categories such as nonabsolutive, conjunction, togetherness, reciprocity, and sameness [Gil (2004b)]. Alongside individual words, the category S includes all multiword expressions in the language, among which are makan ayam and ayam makan in (8). However, in addition to the open syntactic category S, there is also a closed syntactic category S/S, which contains a couple of dozen semantically heterogeneous words, including kalau, marking topics; tiap ‘every’; denggan ‘with’, ‘and’; and others. Thus, Riau Indonesian comes close to being purely monocategorial; it is only the existence of the closed syntactic category S/S that prevents it from actually being so.

The syntax of Riau Indonesian can be stated very simply. Syntactic structures are hierarchic but unordered trees, in which each node is labeled with one of the two syntactic categories, S and S/S. Of course, in any physical representation of such trees on a page, it is impossible not to introduce a linear order; however, it is important to keep in mind that such order is not part of the actual representation. A number of scholars working within different theoretical frameworks have provided arguments in support of unordered tree structures and the representational separation of hierarchic structure and linear order [see, for example, Sanders (1975), Keenan (1978), Keenan and Faltz (1986), Kayne (1994), Bury (2005)].

Syntactic tree structures are formed from subtrees of the following two kinds:

(9) Two kinds of subtrees
In (9a), \( n \) expressions belonging to S combine with each other to constitute a superordinate S. Although there is no strict upper limit on the size of \( n \), branching is most commonly binary, and rarely goes beyond ternary. Formally, (9a) has the structure of a coordination, in which each of the constituent parts is equally ranked. In (9b), a single word belonging to the category S/S combines with a single expression belonging to the category S to yield a superordinate expression of category S. The category name S/S reflects this fact, making use of the familiar “slash” operator from categorial grammar. Unlike (9a), the structure in (9b) is asymmetric. Complex hierarchic structures are built up recursively from the two kinds of subtrees represented in (9).

For present purposes, we shall be concerned only with the former of the two kinds of subtrees. If we set \( n = 2 \), (9a) yields a representation for the syntactic structure of the two sentences in (8):

(10) **Syntactic structure of (8):**

```
       S
      / \      /
     S   S
makan ayam
```

Since the structure in (10) is unordered, it provides an equally appropriate representation for both *makan ayam* in (8a) and *ayam makan* in (8b). To say that the two words *makan* and *ayam* belong to the same syntactic category S is to say that they exhibit the same syntactic behavior, including, specifically, the same distributional privileges as each other and as all other members of S. In addition, as indicated above, they can combine with each other to yield the superordinate S expressions *makan ayam* and *ayam makan*, which, once again, share the same syntactic behavior and distributional privileges. As suggested in (10), the two sentences *makan ayam* and *ayam makan* have an identical syntactic structure, that of a sentential coordination. This reflects the fact, discussed in the previous section, that the two constituent words are equally ranked, lacking in any structural asymmetries such as agreement, government, and the like.

This, then, in a nutshell, is the syntax of Riau Indonesian. So far, I have found no evidence for syntactic categories other than the open category S and the closed category S/S, and no evidence for syntactic structures other than those that can be built up recursively from subtrees such as those in (9). In particular, I have found no evidence for any kinds of empty syntactic positions, or for any kinds of structural dependencies of the type commonly expressed by rules of movement. Thus, on the available evidence, there is indeed ample reason to characterize the syntax of Riau Indonesian as being very simple, in fact close to monocategorial.

### 4.2.2. Semantics

The fundamental semantic structure of Riau Indonesian is also very simple. Every expression in Riau Indonesian has a basic semantic structure in the form of an
unordered tree that is isomorphic to that of its syntactic structure: each node of the semantic structure represents the interpretation of the corresponding node of the syntactic structure of the expression. Whereas the interpretation of terminal nodes is specified in the lexicon, that of nonterminal nodes is derived by compositional principles from that of their constituent nodes.

In fact, most of the compositional semantics of Riau Indonesian can be captured in a single simple rule making reference to the Polyadic Association Operator:

(11) **Polyadic Association Rule of Semantic Compositionality:**

Given a syntactic structure \([X \ X^1 \ ... \ X^n]\) \((n > 1)\) where \(X^1 \ ... \ X^n\) have interpretations \(M^1 \ ... \ M^n\), respectively, \([X \ X^1 \ ... \ X^n]\) is assigned the interpretation \(A(M^1 \ ... \ M^n)\).

The Polyadic Association Rule says, quite simply, that whenever two or more expressions are combined, the meaning of the combination is obtained by applying the Association Operator to the meanings of the individual expressions. In other words, when \(X^1\) to \(X^n\), with meanings \(M^1\) to \(M^n\), are put together, the resulting meaning is \(A(M^1 \ ... \ M^n)\), or ‘entity associated with \(M^1\) to \(M^n\).’ Since the constituent meanings \(M^1\) to \(M^n\) are unordered and equally ranked, the resulting meaning may be characterized as a conjunction. Thus, the Polyadic Association Rule provides a unified semantic representation for Riau Indonesian sentences, reflecting their characterization as vague with respect to thematic roles and ontological types.

The way in which the Polyadic Association Rule works may be illustrated through the semantic representation that it provides for the sentences in (8):

(12) **Semantic structure of (8):**

‘entity associated with eating and with chicken’

\[A(\text{EAT}, \text{CHICKEN})\]

\[\text{EAT} \quad \text{CHICKEN}\]

In (8), *makan* means \(\text{EAT}\) and *ayam* means \(\text{CHICKEN}\). The Polyadic Association Rule applies to the collocation of these two meanings, and assigns them the interpretation \(A(\text{EAT}, \text{CHICKEN})\), ‘entity associated with eating and with chicken.’ Since \(\text{EAT}\) and \(\text{CHICKEN}\) are unordered and equally ranked, the interpretation \(A(\text{EAT}, \text{CHICKEN})\) is thus a completely symmetric conjunction. The above structure constitutes a single unified meaning, encompassing the entire range of interpretations of the sentences in (8), including, among others, those expressed by the various translations of (8) into English provided in Table 1. In particular, it accounts for indeterminacy with respect to thematic roles, allowing for the chicken to assume any role whatsoever in relation to the eating; and for indeterminacy with respect to ontological types, permitting *makan ayam* and *ayam makan* to denote activities, things, places, times, and so on.

The Polyadic Association Rule thus constitutes the basic mechanism governing semantic compositionality in Riau Indonesian. In doing so, it provides a way to
represent the semantic indeterminacy that is so prevalent in the language. The basic semantic structures produced by the Polyadic Association Rule are of an absolutely minimal degree of specificity, adding nothing substantive to the combination of the constituent meanings other than to say that they are related in some way – exactly how is left open to context. The central role that the Polyadic Association Rule plays in the compositional semantics of Riau Indonesian therefore supports the characterization of Riau Indonesian as an associational language.

4.2.3. Further analysis

In the preceding pages, we have seen how Riau Indonesian exhibits each of the three properties of IMA Language to a substantially greater extent than many other languages, and perhaps also to a greater degree than is often supposed to be possible in a natural human language. Nevertheless, Riau Indonesian is still a considerable way off from the limiting case of Pure IMA Language.

The Polyadic Association Rule produces basic semantic structures forming a skeleton which may be subsequently fleshed out by further more specific semantic rules applying whenever appropriate to produce more elaborate representations, involving domains such as coreferentiality, quantifier scope, conjunctive (focus) operators, and many others.

One of the most important kinds of semantic enrichment is that of head-modifier structure. Headedness may apply wherever hierarchical tree structure is present, in accordance with the following rule:

(13) **Headedness Assignment Rule:**

Given a structure \([X \, X_1 \, \ldots \, X_n]\), one of its constituents, \(X_j\), may be coindexed with the entire structure for similarity: \([X \, X_1 \, \ldots \, X_j \, \ldots \, X_n]\). In a structure \(X\) consisting of \(X_1\) to \(X_n\), one of the constituents, \(X'_j\), is singled out as bearing a resemblance to the entire structure, \(X\), with respect to a certain unspecified feature. In such a case, \(X'_j\) is said to be the head of the structure \(X\), and all the other \(X'_i\) (where \(i \neq j\)) are the modifiers of \(X'_j\).

Headedness, as defined above, is present in a variety of cognitive domains [see Gil (1985) for discussion]. Imagine a plate on which, in roughly equal proportions, are a piece of chicken, some beans, and a mound of rice. Whereas Europeans consider the chicken to be the head, and thus conceive of the entire plate as a chicken dish, Southeast Asians consider the rice to be the head, and therefore conceive of the entire plate as a rice dish. Thus, both Europeans and Southeast Asians assign headedness to dishes of food, though the specific choice of head varies across cultures. Of course, headed hierarchic structures are not specific to food; in one guise or another, headedness lies at the heart of many theories of particular domains of cognition. For example, in tonal music, heads and modifiers form the basis of Lerdahl and Jackendoff’s (1983) theory of time-span reductions, which accounts for the ways in which a complex melody may be
successively stripped of its less-important modifying elements, retaining at each stage a smaller and smaller skeletal melody consisting entirely of heads. Within language, too, headedness, as defined in (13) above, is present in a variety of domains, ranging from narrative discourse, as suggested by Shen (1985), through syntax, as in X-bar theory proposed by Jackendoff (1977), all the way to syllable structure, as argued by Anderson and Ewen (1987).

The following rule, a particular case of the Headedness Assignment Rule in (13) above, assigns headedness in the domain of basic semantic structures built up by the Polyadic Association Rule:

(14) **Headedness Assignment Rule for Associative Interpretations:**

Given an associative interpretation \( A (M^1 \ldots M^n) \), one of its constituent substructures, \( M^j \), may be coindexed with the entire semantic structure for coreference: \( A (M^1 \ldots [M^j] \ldots M^n) \).

In (14), the general notion of similarity referred to in (13) is replaced by a more specific kind of similarity, namely, coreference. In a headed semantic structure, the head constituent projects its referential identity up to the entire meaning, whose range of interpretations is accordingly narrowed down. Thus, a headed semantic structure is more specific than the corresponding headless structure.

The effect of assigning headedness to basic semantic structures may be illustrated through the application of headedness to the interpretation of the two sentences in (8) shown in (12) above:

(15) **Semantic structure of (8) enriched with headedness**

\[
\text{‘eating associated with chicken’} \\
[A (EAT, CHICKEN)]_i
\]

\( EAT_i \) \quad \text{CHICKEN}

\( (a) \)

\[
\text{‘chicken associated with eating’} \\
[A (EAT, CHICKEN)]_i
\]

\( EAT \) \quad \text{CHICKEN}_i

\( (b) \)
In (15) above, headedness is depicted twice: by coindexation, in accordance with the definition in (14), and, redundantly but for purposes of greater clarity, by an extra dashed line running up from the head constituent to the root node, tracing the projection of referential identity. The two structures in (15) represent the two possible assignments of headedness: in (15a) EAT is assigned headedness, while in (15b) CHICKEN is the head.

By projecting referential identity, headedness narrows down the range of possible interpretations of the superordinate meaning. In (15a), the head EAT projects its identity up to [A (EAT, CHICKEN)]. Accordingly, the superordinate interpretation no longer denotes an arbitrary ‘entity associated with eating and chicken’ but rather the actual eating, or, more specifically, ‘eating associated with chicken.’ The resulting interpretation retains its indeterminacy with respect to number, definiteness, tense, aspect, and thematic roles, but loses its indeterminacy with regard to ontological type, which is now identical to that of the head EAT, namely, activity. Thus, when EAT is assigned headedness, the interpretation necessarily denotes eating. Some of its possible translations into English might include ‘The chicken is eating,’ ‘Someone is eating the chicken,’ and so forth. Conversely, in (15b), the head CHICKEN projects its identity up to [A (EAT, CHICKEN)]. As before, the superordinate interpretation no longer denotes an arbitrary ‘entity associated with eating and chicken’ but instead denotes the actual chicken, ‘chicken associated with eating.’ Again, the resulting interpretation retains its indeterminacy with respect to number, definiteness, tense, aspect, and thematic roles, while losing its indeterminacy with regard to ontological type, which is now identical to that of the head CHICKEN, namely, thing. Accordingly, when CHICKEN is assigned headedness, the interpretation necessarily denotes chicken. Some of its potential translations into English might include ‘The chicken that is eating,’ ‘The chicken that someone is eating,’ and so forth. Thus, as illustrated above, headedness reduces the range of available interpretations of the superordinate meaning, limiting them to ones that are consistent with properties projected upwards from the head constituent.

How prevalent is headedness in the syntax of Riau Indonesian? As evidenced by the widespread occurrence of vagueness with respect to ontological categories, the Headedness Assignment Rule frequently fails to apply. If the semantic structures of such examples were headed, the head would project its ontological type up to the superordinate interpretation, and, in so doing, restrict its meaning to one ontological type to the exclusion of the other. Nevertheless, in many other cases, the superordinate interpretation is more limited, in ways that suggest that the semantic structure may indeed be headed. So from the availability of both possibilities, it must be concluded that application of the Headedness Assignment Rule is optional.

It should be noted, however, that in real life, that is to say, when analyzing naturalistic data, it is often the case that although two different interpretations are available, one is more prominent, and the other more difficult to obtain. In order to represent such states of affairs, it is necessary to introduce an element of fuzziness into the description. This is easily achieved by attributing to each potentially available assignment of headedness a scalar figure representing the degree to which it is conceptually salient. An essentially identical mechanism, involving preference rules, plays a central role in
Lerdahl and Jackendoff’s (1983) theory of tonal music, accounting for, among other things, assignments of headedness in the form of time-span reductions.

The headedness of semantic structures underlies a variety of syntactic and semantic phenomena in Riau Indonesian. Thus, work in progress [see Gil (2000a)] suggests that semantic headedness may account for certain observable effects involving long-distance dependencies bearing a superficial resemblance to island constraints in other languages. Similarly, as argued in Gil (2005b), semantic headedness may be invoked to account for a large proportion of the word-order constraints that are in evidence in Riau Indonesian. The relevant principle is as follows:

(16) **Head-initial order:**

Heads precede modifiers.

Since headedness is optional, this principle applies only in those cases where headedness is assigned. Nevertheless, head-initial order alone accounts for much of the “word-order typology” of the language.

To see how this works, let us examine the application of head-initial order to (8) above. Some of the apparent properties of the respective constructions, those that might perhaps be attributed to them within conventional grammatical descriptions, are indicated on the right-hand side, within scare quotes.

(17) **Semantic structure of (8) enriched with headedness and linearized**

(a) ‘eating associated with chicken’

\[ \{ \text{A} (\text{EAT}_i, \text{CHICKEN}) \} \]

(b) # ‘eating associated with chicken’

\[ \{ \text{A} (\text{EAT}_i, \text{CHICKEN}) \} \]

\[ \text{makan} \rightarrow \text{ayam} \]
In (17a) and (17b), EA T is the head, as per the semantic structure in (15a), and therefore the expression as a whole denotes eating. In contrast, in (17c) and (17d), CHICKEN is the head, in accordance with the semantic structure in (15b), and hence the expression as a whole denotes chicken. In a conventional grammatical description, (17a) and (17b) would be characterized as having a “sentential” interpretation in which the eating is “predicative,” while (17c) and (17d) would be associated with a “nominal” interpretation in which the eating is “attributive.”

Within each of the above pairs, the preferred head-initial linearization, shown first, is contrasted with the dispreferred head-final linearization, shown beneath it, and marked with a #. Comparing (17a) and (17b), we see that in (17a) head-initial order creates the appearance of a “verb-initial” order, while in (17b) head-final order creates the opposite appearance of a “verb-final” order. Thus, the preference for head-initial order, in accordance with (16), results in an apparent preference for “verb-initial” order, as in sentence (8a), over “verb-final” order, as in sentence (8b). Moving on to (17c) and (17d), we see that in (17c) head-initial order creates the appearance of a “noun-initial” order, while in (17d) head-final order creates the opposite appearance of a “noun-final” order. In this case, then, the preference for head-initial order results in an apparent preference for “noun-initial” order, as in sentence (8b), over “noun-final” order, as in sentence (8a). Thus, if the speaker wishes to convey a specifically “predicative” meaning, with EA T as head, head-initial order will entail a preference for sentence (8a), makan

\[
\begin{align*}
\text{(c)} & \quad \text{‘chicken associated with eating’} \\
& \quad \text{[ A (EAT, CHICKEN)]} \\
& \quad \text{CHICKEN \quad EAT} \\
& \quad \text{ayam \quad makan}
\end{align*}
\]

\[
\begin{align*}
\text{(d) #} & \quad \text{‘chicken associated with eating’} \\
& \quad \text{[ A (EAT, CHICKEN)]} \\
& \quad \text{EAT \quad CHICKEN \quad i} \\
& \quad \text{makan \quad ayam}
\end{align*}
\]
ayam, over sentence (8b), ayam makan. Conversely, if the speaker wants to express a specifically “attributive” meaning, with CHICKEN as head, head-initial order will dictate a preference for sentence (8b) over sentence (8a).

The above observations may be restated from the hearer’s perspective, by taking the sentences in (8a) and (8b) as the starting point. In sentence (8a), makan ‘eat’ precedes ayam ‘chicken.’ Head-initial order, as in (17a), creates the appearance of a “verb-initial” order, whereas head-final order, as in (17d), creates the appearance of a “noun-final” order. Thus, the preference for head-initial order entails a preference for the “verb-initial predicative” interpretation over its “noun-final attributive” counterpart. In contrast, in sentence (8b), ayam ‘chicken’ precedes makan ‘eat.’ Head-initial order, as in (17c), creates the appearance of a “noun-initial” order, whereas head-final order, as in (17b), creates the appearance of a “verb-final” order. Thus, the preference for head-initial order entails a preference for the “noun-initial attributive” interpretation over its “verb-final predicative” counterpart.

Thus, head-initial order, as formulated in (16) above, creates the appearance of a language with an array of word order correlates characteristic of a “verb-initial” language, including, in particular, “noun-attributive” order within NPs. Moreover, it does so on the basis of the grammatical description presented above, making reference solely to syntactic structures involving a single open syntactic category $S$, plus polyadic association and headedness. Thus, even a Relative IMA Language, with an extremely impoverished inventory of grammatical categories, can still appear to display a typologically conventional pattern of word-order preferences.

Still, the above account is not entirely complete. Consider the predictions of head-initial order with regard to the sentences in (8), as represented in (17). Head-initial order makes the correct predication that (17a) and (17c) will be preferred, and also the correct prediction that (17d) will be dispreferred; however, its prediction that (17b) will be dispreferred is in need of further qualification. Let us take a closer look at (17b), representing sentence (8b), ayam makan, with a head-final semantic structure characterizing makan ‘eat’ as head, and therefore denoting eating. In actual fact, the acceptability of (17b) is dependent on the thematic role of ayam ‘chicken.’ For most potential thematic roles, (17b) is indeed dispreferred relative to (17a), in accordance with head-initial order. However, when ayam ‘chicken’ is the agent of makan ‘eat,’ (17b) is more acceptable, and, in fact, is as readily available as its head-initial counterpart in (17a). In other words, in order to say ‘The chicken is eating,’ (8b) ayam makan is every bit as good as (8a) makan ayam: Riau Indonesian appears to be at least as “verb-medial” as it is “verb-initial.” In order to account for such facts, additional principles of linearization making specific reference to thematic roles are thus required. More generally, as argued above, many semantic structures in Riau Indonesian are unheaded, and for them, of course, the principle of head-initial order is inapplicable. Nevertheless, all such structures end up underlying strings of words that occur one after another in nonrandom fashion, thereby revealing the need for additional principles of linearization. Two such principles governing word order in Riau Indonesian, which make reference to iconicity and information flow, are proposed in Gil (2005b).
4.3. Riau Indonesian: A Relative IMA Language

As shown in the previous pages, Riau Indonesian comes closer than is commonly thought possible to displaying the three properties of IMA Language: morphologically isolating, syntactically monocategorial, and semantically associational. In comparison to languages such as Russian, Riau Indonesian bears a closer resemblance to other instances of IMA Language such as the language of pictograms, captive apes, and infants. Of course, Riau Indonesian is still a long way from instantiating Pure IMA Language: it has some morphology, a second closed-class syntactic category, and various construction-specific rules of semantic interpretation. However, as suggested in Section 4.2, above, the bulk of the language is indeed pure IMA: nothing beyond IMA structure is necessary for the morphological, syntactic, and compositional-semantic analysis of basic sentences such as (8). Similarly, more extensive investigations show that nothing beyond IMA structure is required for the representation of the most important properties of sentences whose translational equivalents in English involve complex constructions such as questions, reflexives, relative clauses, and sentential complements. Indeed, examination of the various non-IMA items in Riau Indonesian – the bound morphemes and the members of the closed syntactic category S/S – suggests that they form a heterogeneous set, with no specific characteristic functions of importance to the overall organization of the language.

As a Relative IMA Language, Riau Indonesian is thus simpler than many other natural human languages. In previous eras, there was a widespread belief that, in comparison with their European counterparts, the languages of Africa, Asia, and the Americas were simpler, or more primitive, or plain inferior; in many cases, these assumptions about the languages were coupled with other assumptions about their speakers which today would be judged as morally reprehensible. With the advent of modern linguistics and greater familiarity with the world’s languages, such beliefs were duly discarded; however, their place was taken not by serious empirical investigation of the issues involved, but rather by another dogma, to the effect that all languages are of roughly equal overall complexity. In part, this dogma stems from extraneous considerations having to do with “political correctness”; but there are other, more substantive motivations: linguistics over the course of the last century has simply chosen to concern itself with a different range of issues. Moreover, and perhaps most importantly, complexity of linguistic structure is a notion that is extremely difficult to formalize in an explicit and quantitative manner. None of these factors, however, should be reasons not to try and address the issue of complexity, as indeed is suggested in a number of recent articles, including Comrie (1992), Romaine (1992), and McWhorter (1998, 2000, 2001a, b).

Given the efficiency with which Riau Indonesian fulfills the multifarious functions of a natural human language with so little grammatical machinery, one can only wonder why all languages are not like Riau Indonesian or, for that matter, Pure IMA Languages. One reason for this might be diachronic. McWhorter (1998, 2000, 2001a, b) argues that languages tend to accrue grammatical complexity over the course of time.
In support of this argument, he claims that newly created creole languages are invariably characterized by lesser overall complexity than most “older” languages with their lengthy, continuous, and uninterrupted histories. Accordingly, if a language started off as a Pure IMA Language, grammaticalization processes would soon endow it with morphological structure, syntactic categories, and construction-specific semantic rules, and it would thereby lose its IMA characteristics. So the reason there are no natural Pure IMA Languages may be simply that today’s languages have been around for too long. However, the simplicity of Riau Indonesian shows that the accretion of complexity cannot be construed as an inexorable monotonic process. Riau Indonesian is not a creole and has no known history of radical restructuring of any kind, and yet it is at least as simple in its overall grammatical structure as many creole languages [see Gil (2001a) for detailed argumentation]. Comparison of Riau Indonesian with related Austronesian languages suggests that their common ancestor, Proto-Austronesian, spoken perhaps 5000 years ago, was substantially more complex than Riau Indonesian in many grammatical domains. Thus, at some stage between Proto-Austronesian and Riau Indonesian, the accretion of complexity must have been reversed, in order for Riau Indonesian to emerge, gradually over time, as a Relative IMA Language. Given that this happened at least once, the question arises once again of why more languages could not have taken the same path, and indeed of why some languages could not have gone further along the path of simplification to end up as Pure IMA Languages. I have no answer to this question. Perhaps the diachronic forces that produce complexity are just “stronger” than those that operate in the opposite direction toward simplicity. Or perhaps the greater complexity of most other languages is due to entirely different factors.

5. Cognition

The presence of IMA Language in artificial semiotic systems, in early child language, and to a considerable degree in some natural languages, suggests that IMA Language is an important feature of human cognition. However, to see how this is so, we need to adopt a somewhat different way of looking at IMA Language. The definition in (1) and much of the subsequent discussion were of an essentially negative nature; that is to say, IMA Language was taken to be language without particular features: morphological structure, syntactic categories, and construction-specific semantic rules. This choice of perspective reflected our presuppositions regarding what languages are like: we expect them to possess these particular features, and so it requires a mental effort to entertain the possibility of languages lacking them. However, in order to appreciate the role of IMA Language in human cognition, we shall adopt an alternative positive perspective, focusing instead on particular features that IMA Language does have.

Three such features are the following:

(18) (a) Recursive Tree Structure
     (b) Sign-Meaning Pairings
     (c) Associational Compositional Semantics
While all IMA Languages possess the above features, they do not define IMA Language, since they are also exhibited by more complex languages. On the other hand, Bickerton’s (1990) protolanguage fails to exhibit the first property, syntactic recursion, though it shares the latter two with IMA Language. None of the three features is specific to natural language, and therefore none should be considered to be part of a domain-specific Universal Grammar. Rather, each of these features is manifest in a variety of cognitive domains, and may therefore be attributed to general human cognition.

Recursive tree structure refers to the ubiquitous human cognitive ability to construct groupings. Presented with a collection of objects, we view them as clustering together into groups according to various criteria: spatial or temporal organization, size, shape, color, quality, etc. The groups then constitute new objects which, in turn, cluster into larger groups, and so on over and over again. The cognitive ability to form groupings has been the focus of a considerable amount of investigation; in particular, this ability is essential to all higher-level cognitive capacities, as shown for example in Lerdahl and Jackendoff’s (1983) theory of tonal music. It is this general cognitive ability that is also manifest in the tree structures that represent the syntactic structures of IMA Language, as presented in an array of pictograms, a sequence of signs by a captive ape, an utterance by a young child acquiring English, or a sentence in Riau Indonesian.

Sign-meaning pairings reflect another general human cognitive ability, one that lies at the heart of semiotic theory. Although the quintessential realization of sign-meaning pairings is that evident in the lexicons of natural languages, similar pairings occur in many other domains: consider a commercial logo such as that for Apple computers, a red traffic light, or the individual pictograms in examples such as that of the arrow and bicycle. In semiotic theory, a major concern is to demonstrate that sign-meaning pairings are everywhere around us; in so doing, the notion of conventionalized sign-meaning pairing is sometimes extended well beyond its central domain of applicability, and consequently watered down considerably, sometimes to the point of vacuousness. Nevertheless, even under a conservative construal of the notion of sign-meaning pairing, it is clear that this feature of IMA Language too is not specific to language but rather part of general human cognition.

Associational compositional semantics pertains to the cognitive ability to assign interpretations to combinations of signs in accordance with the Polyadic Association Operator, as defined in (2) above. Again, this feature of IMA Language is clearly of a general, nondomain-specific nature. Whenever we encounter a collocation of two or more sign-meaning pairings, we assume that the collocation is intentional, and attribute to it a meaning that has to do in some way with the meanings of the individual constituent signs. Imagine the rather unlikely juxtaposition of an Apple computer logo with a red traffic light. When encountering such a novel combination, we assign it the meaning A (APPLE.COMPUTERS, STOP), ‘entity associated with Apple computers and with stopping,’ in accordance with the Polyadic Association Operator, and then seek for some contextually plausible interpretation, perhaps ‘Stop here for Apple computers,’ or alternatively ‘Don’t use Apple computers.’ A substantial recent literature in cognitive psychology may be construed as dealing with the mechanisms whereby the broad meanings...
assigned by the Polyadic Association Rule are narrowed down in particular contexts [see, for example, Murphy (1990), Estes and Glucksberg (2000), Wisniewski (2000)].

Thus, IMA Language, with its three properties presented in (18), is a characteristic feature of general human cognition. In its most transparent form, IMA Language is evident in semiotic systems such as the language of pictograms. It may also be considered as a foundation on which the more elaborate and domain-specific structures of natural human languages are constructed. Moreover, just as houses are built from the foundations up, so children acquire natural languages beginning with IMA Language. Our prehuman ancestors also began with IMA language, thereby bequeathing it to us as an evolutionary relic from our distant past, albeit one that is very much alive and with us in the present.

Acknowledgments

I would like to thank the many speakers of Riau Indonesian who gave me a new perspective on language, and my many colleagues at the Max Planck Institute for Evolutionary Anthropology who provided the intellectual environment that made it possible to develop that perspective into the ideas presented in this chapter.

References


This page intentionally left blank
Abstract

Our goal in this chapter is twofold. First, we describe some basic properties of categorization in Langue des Signes Québécoise (Quebec Sign Language: LSQ). Second, properties of the gestural modality of a sign language seem to bring about a categorization different from that induced by the modality of an oral language; therefore, we explore some consequences for the categorization of linguistic material in general and the nature of “universality” in linguistics.
1. The categories of lexical items

One method of analysis has dominated western grammar: the units of language are assigned to categories. These categories are described by sets of features that determine the co-occurrence possibilities of the units [see, e.g., Travis (this volume)]. This method is based on an old assumption that it is easier to describe phenomena if we assume that a term like *destruction* belongs to the same syntactic category as the term *dog*. This kind of approach can already be found in the work of Varro and Apollonius (2nd century) and Priscian (5th century). It raises several questions, which we will only briefly review here.

First, what are the criteria for any particular category? The standard answer is that there are three kinds of converging criteria (already in Varro’s *De lingua latina*):

1. a. morphophonological: common inflections, derivational marking
   b. syntactic: an element of category A combines with one of category B
   c. semantic: differences in the way of signifying (Modists in Middle Ages)

A problem with these criteria is that, on the surface, some languages have very few morphophonological markings, and compensate with a very rigid linear order in syntax, whereas others have very free syntax and a lot of morphophonological markings¹. More directly related to our concerns is the fact that these criteria are quite intimately tied to the modality of oral languages. How valid are they for sign languages?

A second question raised by categorization concerns is the problem of what is actually stored in the mental lexicon. There are two main polarizations here:

2. a. Categorized forms, which are selected according to their category;
   b. Noncategorized forms, where categorization expresses the particular function
   the item holds in a particular use.

The first view is quite standard, but the second one is proposed to account for the categorization of words like English *up*, which seems to fit into several categories, as illustrated in (3).

¹ As for the semantic criterion, it is notoriously difficult to use. The Modists tried to show that there is a correlation between grammatical categories like noun and verb, and ways of signifying (cf. *dolor* (“pain”) and *doleo* (“I suffer”)). The Port-Royal grammarians used the criterion to distinguish nouns (substances) from adjectives (properties). However, they pointed out problems with French words such as the noun *blancheur* (“whiteness”), which expresses a property that has an autonomous existence, and *humain* (“human”) used as an adjective, where we find the opposite situation of an adjective expressing a substance. Despite the difficulties encountered with this semantic criterion, it has such an intuitive appeal that it endures. For instance, Jackendoff (2002, pp. 257–258) maintains the noun–verb distinction mainly on the basis of a similar distinction in the way of signifying: “Nouns can express any semantic category whatsoever: not just objects but situations (*situation, concert, earthquake, perusal*), times (*Tuesday, millennium*), and so on. But verbs can express only situations (events, actions, and states).” In Generative Grammar in general, following the structuralist tradition in this case, the morphophonological and syntactic criteria have been more widely used because they are more easily formalized. However, the casual intuition behind the traditional semantic criterion has underlain categorization from the outset: “The question of substantive representation in the case of the grammatical formatives and the category symbols is, in effect, the traditional question of universal grammar. I shall assume that these elements too are selected from a fixed, universal vocabulary …”

(Chomsky (1965, pp. 65–66)
The view that lexical items are not marked for grammatical category has been given much exposure in recent years in the Distributed Morphology framework [Halle and Marantz (1993), Marantz (1997), Harley and Noyer (1999), Barner and Bale (2002)]. For instance, in such a grammar, lexical roots like \textit{grow} and \textit{destroy} are category-neutral: “when placed in a nominal environment the result is a ‘nominalization’; when the roots are placed in a verbal environment they become verbs” [Marantz (1997)].

The question then is, of course, what is a “nominal environment” or a “verbal environment”? Distributed Morphology provides a formal answer to this question. It assumes that, like phrases, words are built in the syntax. Roots are inserted into syntax, where functional heads determine their status as nouns or verbs. In turn, phonological forms are inserted based on the featural status of roots in context. However, this is not a very satisfactory answer, as it complicates the system without really answering the question; it merely postpones it.

First, the categorial distinctions are not dispensed with, but actually duplicated: the lexical items themselves are marked as being selected by certain functional/categorial heads; in addition, the functional heads are names of constructions in a syntactic environment. The duplication is clear if we compare this with a similar approach already found in Chomsky’s “Remarks on Nominalization” (1970, p. 21), which did without these functional heads.

Moreover, special lexical rules such as conversion are replaced by special phonological rules of late insertion: this is not much of a qualitative improvement. Worse, the system has a problem of compatibility with Bare Phrase Structure (BPS). As Barner and Bale (2002) acknowledge, this view requires “a syntactic system that allows for complex heads in the phrase structure (consisting of the root plus the functional head).” They give two candidates: \texttt{[X-head R F]} and \texttt{[X-head-1 R [X-head-2 F]]}, which they say are compatible with BPS. Not so: labels have no status in BPS, whereas here they seem to have one.

This model basically says that when an item appears in a certain syntactic (and semantic) context, it takes form \texttt{Y}. There is no need of a functional category (FC) for this: for each functional head identifying a category, it is necessary to determine in what syntactic (and semantic) context it may appear; this is sufficient to “identify” the category of the item. There is no need of a label or functional head: this just postpones the answer to the question of categorization: criteria are then needed to determine the category of the functional head. So why not apply them directly to the lexical items in context? Under this view, classical categories are a bit like the notions of subject or object: they are convenient names for a relation, but have no status in the theory itself.

Syntax is the theory of the relational properties. Once these properties are properly identified by the theory, categorial labels may also be merely convenient names for a relation. A theory with functional heads that determine categorial status produces a hybrid system in which some elements are noncategorized forms – the lexical roots – and other elements are categorized forms – the functional heads. The latter are redundant and
therefore should not be maintained in the grammar. Moreover, by dispensing with these categorization heads, we avoid a pitfall of functional categories: they quickly lead to taxonomic theories, to a grammar of lists. Thus, if we look at additions made to the traditional inventory in the last two decades, many new “functional” categories have been introduced on the basis of casual intuitions similar to those behind the traditional semantic criterion for categorization. For instance, Cinque (1994) proposes the categories Quality, Size, Shape, Color, Nation, and Speaker-oriented (see Section 2.2). Beghelli and Stowell (1997, pp. 74–75) suggest WhP, NQP (negation), Distributive Phrase, Ref Phrase, and Share Phrase (“interpreted with ‘dependent’ specific reference”). Munaro and Obenauer (2000) introduce EvCP (Evaluative CP = a kind of “Opinion Phrase”). Munaro et al. (2001) assume the categories Interrogative Force, Focus, Operator, and Topic. This use of functional categories that extend to discourse notions, like various illocutionary forces, and pragmatic notions, such as speaker attitudes, suffers from the same general weakness as Generative Semantics: the reliance on casual intuitions is too unconstrained and too vague, as indicated in Bouchard (2002, p. 334). The common characteristic of these analyses is the methodology of listing positions by means of functional categories and listing placements of constituents by means of uninterpretable features. But lists are merely assertions of existence: they are inventories of facts, they tell us what is. Lists do not deal with modalities of existence: they do not tell us what is possible. In the current absence of constraints on possible uninterpretable categories or features, a list like the nonsubstantive part of the lexicon reveals nothing about what variations are possible in languages or crosslinguistically, nor what the limits on variation are. In short, no deductive science is possible under these conditions. Categorization heads fall into this general pattern of listing: they constitute a list of convenient names for the relations that occur between items, but they provide no indication about why these should be the relations that are possible in language.

In addition to this foundational problem with lists of categories, there is also a methodological difficulty in comparative grammar that Humboldt had already raised: we must be careful in applying to an unknown language the categories of a known language that express a similar notion. We would add that this is especially true when the physiological means of combining elements are different, as in sign languages. In order to gain a better understanding of how modality may affect categorization, at least in part, we now turn to basic categorization data in Langue des Signes Québécoise (LSQ).

2. Traditional categorization applied to LSQ

In sign languages, numerous noun–verb pairs that are related by meaning also have the same formal features (configuration, orientation, place of articulation, and contour of movement). In early studies on American Sign Language (ASL), this led scholars to conclude that noun–verb pairs are phonologically identical [Stokoe, Casterline and Croneberg

2 Often, the functional categories and uninterpretable features are ad hoc to the point that they are not even provided with particular categorial identification – they are just placeholders labeled as FP or triggers labeled F.
(1965/1976). However, others have remarked that noun–verb pairs are phonologically distinct [Supalla and Newport (1978)]: for nouns, the movement is repeated and reduced, whereas for verbs, the movement is long, possibly repeated, and continuous or held. In LSQ, some semantically related noun–verb pairs are phonologically distinct in this way:

(4) a. CHAIR/TO-SIT
    b. CAMERA/TO-FILM
    c. AIRPLANE/TO-FLY [in an airplane]

However, most noun–verb pairs are phonologically identical:

(5) a. HELP/TO-HELP
    b. TEACHING/TO-TEACH
    c. INTERPRETER/TO-INTERPRET

A noun–verb categorial distinction cannot be made for these forms in isolation. It is only in the context of their use that they can be distinguished, as in (6).

(6) a. TO-INTERPRET/INTERPRETER [identical forms]
    b. YESTERDAY PIERRE INTERPRET ALL-DAY
       ‘Yesterday, Pierre interpreted all day.’
    c. YESTERDAY INTERPRET INDEX BE VERY-SICK
       ‘Yesterday, the interpreter was very sick.’

Phonologically identical pairs are also frequently found for other categories, and here too the only way to determine the category of an element is by its context of use:

(7) Phonologically identical pairs of nouns and adjectives
    a. COMPETENCE/COMPETENT [identical forms]
    b. FOR WORK TEACHER MUST HAVE COMPETENCE
       ‘For a teaching job, you must have competence.’
    c. TEACHER COMPETENCE INDEX GO
       ‘The competent teacher is leaving.’

(8) Phonologically identical pairs of verbs and prepositions
    a. TO-CONTAIN/IN [identical forms]
    b. COURSE LSQ INDEX IN WHAT
       ‘What does the LSQ course contain?’
    c. MANY CAKE INDEX HAVE NUT IN
       ‘Many cakes have nuts in (them).’

(9) Phonologically identical pairs of adjectives and adverbs
    a. ORAL/ORALLY [identical forms]

---

3 This phonological distinction does not hold for all such forms in a discourse context. It is more regular for items produced in isolation, and even more so when they are produced in noun-verb pair contexts [Johnston and Schembri (1999)].
b. FRENCH LANGUAGE ORAL BELONG QUEBEC
   ‘In Quebec, French is the oral language.’

c. TEACHER UQAM INDEX TEACH ORAL
   ‘All UQAM professors teach me orally.’

Pronouns and definite determiners also have identical phonological forms. In a sign language, a noun is introduced by producing a sign that corresponds to it. Sign languages make extensive use of space to refer to individuals within a discourse. So typically, a noun corresponding to a third-person referent is assigned a locus in the signing space by a pointer (glossed as INDEX). It is this pointer that functions either as a definite determiner or a pronoun, depending on the context of use. The pointer can take the form of signing the noun in that location (10a), pointing at the locus with the index finger (10b), or directing the gaze at the locus, or inclining the body toward the locus.

(10) a. PEN\textsubscript{(loc)} MARY TAKE
   ‘Mary takes the pen.’

   b. PEN INDEX MARY TAKE
   ‘Mary takes the pen.’

When a pointer is used to initially assign a locus to a noun in this way, it functions as a determiner. The assignment of a locus makes the noun definite (11a), whereas if the noun is not localized, as in (11b), it is indeterminate:

(11) a. CAT INDEX I-WANT
   ‘I want the cat.’

   b. CAT I-WANT
   ‘I want the cat/a cat/the cats/cats/cat.’

Once a referent is established at a location in the signing space, a pointer directed toward that location is interpreted as referring back to that specific referent, in which case we could say that the pointer functions like a pronoun. For instance, a signer can reuse some loci by signing a verb from one locus to another to express distinct grammatical functions, as in (12), in which the verb is signed from the locus of LAWYER to the locus of JUDGE. A locus can also be reused by pointing to it as the verb is signed, as in (13), in which the signer first points to the locus of GIRL, and then to the locus of BOY.

(12) JUDGE LAWYER GO-EXPLAIN
   ‘The lawyer goes to explain something to the judge.’

(13) BOY INDEX\textsubscript{a} GIRL INDEX\textsubscript{b} LOVE INDEX\textsubscript{b} INDEX\textsubscript{a}
   ‘The girl loves the boy.’

4 This seems to be a general property of sign languages. Interestingly, it has long been observed that pronouns and definite determiners also have identical phonological forms in many oral languages. This is too common to be accidental. Bouchard (2002, ch. 4) gives a detailed analysis of French le/la/les and argues that the reason why these forms function both as pronouns and as definite determiners is that both categories perform the same function, but with respect to different elements, i.e., a tensed verb and a noun, respectively.
A generalization emerges from the comparison of the various phonologically identical but functionally distinct pairs compared above: the category of an element is usually determined not by its phonological form, but by its function in a particular context of use. We illustrate these differences in the way elements function by looking more closely at two pairs: noun–verb and pronoun-determiner.

2.1. Nouns and verbs

Nouns and verbs generally behave identically in space, but they function differently. Four important characteristics distinguish them.

(i) Nouns generally identify a locus, verbs relate to the established loci of their arguments. This is what we just saw in the discussion of examples (10) – (13).

(ii) Negation applies only to verbs (verb phrases), not to nouns alone:

(14) a. *JEAN\(^{(neg)}\) COOK NOT-KNOW INDEX
b. JEAN COOK NOT-KNOW\(^{(neg)}\) INDEX\(^{(neg)}\)

‘The cook does not know Jean.’

(iii) A possessive marker may be used with a noun, but not a verb:

(15) a. BOSS POSS LAUGH
‘His boss is laughing.’
b. BOSS LAUGH POSS
* ‘It’s the boss laughing.’
‘It’s the boss’s laugh.’

(iv) Aspect applies only to verbs, not to nouns:

(16) a. STUDENT WORK TO-WRITE-LONG
‘Students’ work is long.’
b. STUDENT TO-WORK-LONG
‘The student worked for a long time.’

2.2. Pronouns and definite determiners

Though they have identical phonological forms, pronouns and definite determiners are distinguished by the way they function:

(i) As we saw above, determiners assign a locus, and pronouns reuse a locus.

(ii) Determiners express number, pronouns do not agree in number.

The distinction between a collective reading and an atomic reading – between plural and singular – is made at the initial assignment of a locus: if the pointer is directed at a

\(^5\) In (14), (neg) indicates that a headshake occurs simultaneously to the sign to which it is attached.
point in the locus, the noun gets an atomic/singular reading; if the pointer traces a circular movement, indicating a zone in the locus, the noun gets a collective/plural reading.

(17) STUDENT INDEX(singular) BOOK 3-GIVE-1
   ‘The student gives me a book.’

(18) STUDENT INDEX(plural) BOOK 6-GIVE-1
   ‘The students give me a book.’

However, a locus initially established as a plural may be reused with a singular pointer:

(19) CATa INDEX(plural) MARYb LOVE INDEXb(singular)-INDEXa (singular)
   ‘Mary loves the cats.’

Pointers used to refer back to a previously established referent, i.e., used as pronouns, never agree in number. This absence of number agreement appears to be the case in sign languages in general. This is an interesting and revealing property: it suggests that the presence of number agreement in oral languages and its absence in sign languages is somehow tied to distinctions between the oral and gestural modalities. The correlation is strengthened by the fact that there are actually quite a few differences between pronouns in oral languages and in sign languages.

3. Pronouns in oral languages and in sign languages

Certain features are typically found in the pronominal systems of oral languages. Here is a representative list of properties for which there are often particular pronominal forms [adapted from Givón (1984, p. 354)]:

(20) a. Participant deixis: speaker, hearer, nonparticipant
    b. Inclusion/exclusion of the speaker or hearer
    c. Spatial deixis: proximity, distance, visibility (for third person)
    d. Number: singular, dual, plural
    e. Class/gender
    f. Case markings

Interestingly, in LSQ (and other sign languages), pronouns do not exhibit these standard properties of oral languages. The first three properties are expressed in sign

---

6 There are other ways to express a zone. For example, the index finger can make a linear movement (i), or there can be a series of index pointings (ii), or again several nouns may be signed on the same locus, like MARY, PAUL, and JOHN in (iii):

(i) DOOR INDEX(linear movement) MAN CLOSE CLOSE CLOSE
   ‘A man closes all the doors.’

(ii) SHOE INDEX ++ (series of index pointings) MARY 3-GIVE-1
   ‘Mary gave me some shoes.’

(iii) PAULx JOHNx MARYx (on the same zone) BOOK 1-GIVE-6
   ‘I give a book to Paul, John, and Mary.’
languages, but there are no particular pronominal forms to do this. As for the last three properties, they are wholly absent from pronouns in sign languages. Since there is such a clear-cut modality based demarcation between languages that do and do not have this cluster of properties, the most likely place to find an explanation for the split is in the different physiological substances that the oral and gestural modalities use.

3.1. The effects of perceptual substances on linguistic forms

In language, elements of two orders – the perceptual substance and the conceptual substance – combine [Saussure (1916)]. Since it stems from the combination of elements from these two substances, a linguistic form is determined by them. This makes it possible to establish the basic notions of linguistics by following a method used in other sciences: some propositions are considered to be generally valid – axiomatic – because they are logically prior to linguistic theory, in the sense that they must antedate linguistic theory since its object of study presupposes them. This is true of propositions concerning the two substances that combine to generate a linguistic form: the conceptual substance has its own properties, which are dependent on the structure of the human brain. The brain in which the language system is actualized resides in a human body that has a particular sensorimotor system that determines the properties of the perceptual substance. If different sensorimotor systems are used, as is the case with oral and sign languages, these systems determine perceptual substances with different properties, and these perceptual substances determine different linguistic forms. The different perceptual substances underlying oral languages and sign languages have two principal effects on linguistic forms [see Bouchard and Dubuisson (1995), Bouchard (1996)].

First, the perceptual substances affect the possibility of simultaneity in the languages. For instance, the articulatory system used in oral languages does not allow two words to be produced simultaneously. This is due to the physics of articulation and is explained by biological science. As far as linguistics is concerned, it is an unquestioned observation. When Tesnière (1959) [following Saussure (1916, p. 170)] and Kayne (1994) argue that many properties of syntactic hierarchical structure derive from the fact that words occur in an irreversible, asymmetric temporal sequence, they take this fact for granted and assume that its explanation must be logically prior to the syntactic properties they are studying. On the other hand, the perceptual substance of a sign language does not impose such strong constraints of linearity, in particular because there are several independent articulators: both arms and hands, the whole body (its orientation, inclination), and the head (its orientation, inclination, direction of gaze, facial expressions). Consequently, while simultaneous expressions are quite restricted in oral languages (modulations of intonation superimposed on a sentence to express statements, questions, or exclamations, or intonations on a constituent to express grammatical functions, as in some tone languages), they are very frequent in sign languages. For example,
before the other; in ASL and other sign languages, what may be referred to, with considerable latitude, as “the manual verb” and “the nonmanual adverb” are visible at the same time. Like spoken verbs and adverbs, they can be separated, but separation must be a spatial, not a temporal, operation.

Armstrong, Stokoe and Wilcox (1995, p. 86)

The second effect of perceptual substances on linguistic forms concerns the dimensions involved. The auditory-oral substance is restricted to the single physical dimension of time, whereas the visual-gestural substance involves more dimensions: in addition to the temporal linearity that it shares with the auditory-oral substance, it also involves the three dimensions of space. An effect of this difference in dimensionality is that sound, which is limited to the dimension of time, is ephemeral, whereas signs such as spatial loci have a certain permanency: thus, they may remain indexed and be reused in a discourse.

3.2. Explaining the different properties

With this much background, we can now account for the split between oral and sign languages concerning the properties of pronouns listed in (20). We must determine (i) what contributions these properties make to grammar, (ii) how the perceptual substance of sign languages allows a signer to express the first three properties – participant deixis, inclusion/exclusion, and spatial deixis – without having recourse to special forms, and (iii) why pronouns in sign languages do not need the contributions made by number, class/gender and Case markings.

3.2.1. Participant deixis

The distinction between speaker, hearer, and nonparticipant is a very useful one from a general cognitive perspective, and this is likely the reason why it is expressed in all languages. In an oral language, any possible phonic realization may be connected with the concepts SPEAKER or HEARER. This follows from the general observations of Saussure (1916, pp. 105–110): nothing prevents any particular signifié from being associated with any signifiant to form a sign, because the phonic substance and the conceptual substance provide no logical or natural reason for a language to make any particular association. For instance, there is no inherent reason why a subject SPEAKER is I in English and je in French. Nothing in the concept SPEAKER or in the phonic realizations I and je induces these associations: they are arbitrary conventions in these languages. In other words, the perceptual substance of oral languages is such that arbitrary lexical forms are required to express distinctions such as speaker, hearer, and nonparticipant.

In a sign language, there are no specific lexical forms for these distinctions. Since the substance is spatiotemporal, the body of the speaker and the body of the hearer are part of it and can be directly used as signifiants for the signifiés SPEAKER and HEARER: thus, there is a natural reason for a sign language to make these particular
associations. The nonparticipant is dealt with in a manner that is more similar to what takes place in oral languages: the pointer is directed at a particular locus in the signing space that has been arbitrarily assigned to a noun/actant earlier in the exchange between the signers. However, this arbitrary assignment differs in an important way from what happens in oral languages: the spatiotemporal substance, i.e., the locus, is not assigned to a pro-form, but to a noun/actant that remains actualizable and can be reused by pointing at it. So this is not a case of associating perceptual and conceptual substances anew to produce a sign, as in oral languages, but of associating conceptual substance with “old” perceptual substance, which has a certain permanency in the signing space. Therefore, here too there is a natural reason to make this particular association, namely, the permanency of the locus in the spatiotemporal substance.

3.2.2. Inclusion/exclusion of the speaker or hearer

As in the previous case, there is nothing in the substance of any possible phonetic realization that provides a logical or natural reason for it to express inclusion or exclusion of the speaker or hearer, so arbitrary lexical forms are required. On the other hand, in a sign language, there are no specific items for these distinctions since they can be directly expressed by means of the body of the speaker and the body of the hearer, which are part of the spatiotemporal substance.

3.2.3. Spatial deixis: proximity, distance, visibility (for third person)

The reason why specific lexical items are required to express these notions in oral languages is the same as in the two previous cases. Here again, there are no particular items for these distinctions in sign languages since they can be directly expressed by spatial means. This is clear in the case of distances from the body of the speaker and the body of the hearer. For nonparticipants, the distance can be expressed with respect to an established locus, or by taking advantage of the multiple articulators. In the latter case, two articulators, typically the two hands, are localized with respect to each other; alternatively, in role-playing, the signer’s whole body may represent one actant and one hand can express the distance between it and another actant.

3.2.4. Number

An interesting and revealing property of sign languages is that there is never any number agreement. As we have seen in (17) and (18), when a locus is initially assigned to a referent by a determiner, it is established either as a point (singular) or as a zone (plural). This is a kind of inflection. Yet when a locus is reused, it is always with a point

---

7 The discussion in this section expands on the proposals made by Parisot (2003) and Bouchard and Parisot (2003).
marking, never a zone marking, even though the locus will retain its plural interpretation if it was originally assigned a zone marking. This lack of agreement is not an accidental property: it derives from the composition of the visual-gestural substance.

To understand why there is no number agreement in sign languages whereas pronouns in oral languages generally agree in number with their antecedent, we must understand the place of agreement in the linguistic system. Agreement is a constraint on the features of an element imposed by the features of another element. This constraint does not require that the material forms be identical, but that they express identical values of abstract features in a paradigmatic domain.

In oral languages, agreement is required when we want to say something additional about a referent because these languages are highly restricted to temporal chains of elements that have no permanency. All signifiants are ephemeral: once pronounced, they cannot be reused. For instance, once we have pronounced *The Count* in a sentence like (21), we cannot reuse that signifiant, that physical object: if we want to say something else about the Count, we need a new signifiant.

(21) *The Count left at five o’clock.*

We must either utter a new token of the phrase *The Count*, or introduce a new signifiant that reactualizes the information conveyed by *The Count*, such as the pronouns *he* and *him*. So there is a conflict between the newness of the phonic realization and the anteriority of the referent. The perceptual substance is such that there must be a new signifiant; yet at the same time there must also be some indication of permanency, of a recall of a previously established actualization of a referent. It is the paradigmatic domain that introduces this permanency: membership of a paradigm of abstract features is a permanent property of a word (as far as synchronic grammar is concerned). So in our example, the pronouns *he* and *him* can reactualize the Count because they share some distinctive features in a paradigm with the phrase *The Count*, i.e., [+MASC; +SING; 3PERS]. Since these features can correspond to several elements, additional discourse strategies narrow the possibilities by making one element more salient at a given moment in a given speech situation.

The situation is quite different in sign languages. The visual-gestural modality inherently has such permanency through its use of space: the initial assignment of a locus is permanent (for as long as the signing exchange lasts), so that the locus can be used to reactualize a discourse element and also to indicate the relation that this element holds with others (by starting or ending a verb on the locus, for example). Since a referent can be reactualized by a perceptual substance that has some permanency, i.e., a locus in space, there is no conflict between the newness of the perceptual realization and the anteriority of the referent, and therefore number agreement is not necessary. That is why when a referent is reactualized through the locus initially assigned to it, there is no

---

8 A locus may be reused to indicate different types of zones, but each zone-type expresses an additional distinction of quantification, such as distributivity or partitivity; these zones never express a simple plural.
need to indicate the singular/plural distinction and the pointing is always to a point, not a zone.\footnote{Since there is an unlimited number of locations in space, Lillo-Martín and Klima (1990) conclude that there is a potentially infinite number of distinct pronominal forms in a sign language. The foregoing discussion suggests another conclusion: that there are no pronominal forms in sign language.}

The paradigmatic membership is a subcomponent of Saussure’s \textit{rapports associatifs}, which he opposes to \textit{rapports syntagmatiques}. For instance, \textit{seventy-nine} is related associatively to \textit{seventy-eight, sixty-nine}, etc., and syntagmatically to its elements seventy and nine.

Words which have something in common are associated in the mind […] These links are of a totally different nature from syntagmatic relations. They do not rely on an expanse; their base is in the brain […] The syntagmatic relation is \textit{in praesentia}; it resides in terms which are all present in a series. On the contrary, the associative relation links terms \textit{in absentia} in a virtual, mnemonic series. \textit{Saussure} (1916, pp. 170–171 [our translation])

An associative relation such as number agreement is not necessary to provide a virtual presence for a referent in sign languages because the spatial locus assigned to the referent already has a certain permanency – it remains \textit{in praesentia}. In sum, reactualization of a referent is a linguistic universal. However, the frequency of abstract features (i.e., agreement) as a means of reactualization seems to derive incidentally from the fleeting nature of the acoustic substance of oral languages.

3.2.5. Class/gender

A third-person pronoun may potentially reactualize innumerable nonpresent referents, and class/gender markings help to narrow down the possibilities. We have an indication that this is the motivation for class/gender markings in the fact that they usually appear only on third-person pronouns, and not on first- or second-person pronouns in oral languages: first- and second-person pronouns provide sufficient information to identify the proper referent, because their referents are usually human and face to face, with their spatial deixis and class/gender well established from the context.

The absence of class/gender markings in sign languages can be correlated with the general absence of gender on first- and second-person pronouns in oral languages. A referent is reactualized by pointing to a locus that was previously assigned to it. Since the initial assignment of a locus lasts for as long as the signing exchange, reusing this locus provides sufficient information to identify the proper third-person referent and no additional means need to be added to the grammar.

So, as in the case of the absence of number agreement, the absence of class/gender markings in sign languages derives from the composition of the visual-gestural substance. Its spatial properties ensure a certain permanency of locus assignment, so the identity of proper referents remains very salient. In contrast, the substance of oral languages has no permanency: a new phonic realization has no inherent link with a
previously actualized referent, and class/gender markings help limit the numerous possibilities.

3.2.6. Case markings

There are no morphological markings for Case in sign languages, either on nouns or on any pro-like forms. To understand why this is so, we must determine the role of Case markings in the grammar of oral languages, how they relate to properties of the substance of oral languages, and what difference in the substance of sign languages makes them unnecessary.

Morphological Case markings are a means of indicating what grammatical function a phrase has in a sentence. It has long been observed that there is a trade-off between the diverse means used to encode this information. For instance, as Meillet (1949, 1950), Keenan (1978), and many others have observed, the richer the Case-marking system is in a language, the less rigid is the order of major constituents, and vice versa. This is because Case markings and order are two means which can equally appropriately encode information about grammatical functions.

There are two broad strategies to account for this kind of functional covariation. The first one relies mainly on formal properties: it provides a formal representation of order and Case, assumes for reasons of parsimony of the theory that one of the two is more basic, and accounts for the variation across languages by proposing formal operations which can derive the secondary means from the more basic one. The second strategy derives the variation from properties of the perceptual substance: this substance provides equally valid means to encode grammatical functions, and each language chooses among these means.

Consider the analysis based on formal properties. The currently dominant proposal is that the primary means of encoding grammatical functions is by fixed positions in a phrasal structure, based on the intuition that there are natural positions for the interpretation of major constituents. Deviations from this structural encoding, such as the use of Case markings, are assumed to be superficial: at some deeper level, all relations between predicates and arguments in all languages are encoded by fixed positions in a phrasal structure. There is a universal list of positions that are used to express certain grammatical functions such as subject and object. Major constituents can appear fairly freely in several different positions in Case-marking languages because these languages have special displacement strategies. For instance, Chomsky (2000, p. 145) assumes that Case-marked elements happen to have “a scrambling feature [which] induces pied-piping even after Case assignment, with the pied-piped element ‘attracted’ by a higher probe.” What distinguishes English from Latin, in this view, is

---

10 The traditional view (influenced by studies of Latin) that functional case markers express grammatical functions is revived in Neeleman and Weerman (1999), but it is marginal in mainstream generative grammar. Quite patently, the means of encoding that is assumed to be primary is whichever one prevails in the language which, for some reason, in the theorists’ view, is assumed to be dominant.
that English does not have such “scrambling” features. A language has “scrambling” features whenever it happens to have constituents that are overtly marked for Case. Why this correlation holds is a mystery. A Case-marking language just happens to have extra mechanisms that conspire to give the impression of a freer order. To say that some languages have a list of uninterpretable scrambling features while others do not is just a restatement of the facts. This approach does not say anything about why these choices exist. Moreover, contrary to the claims, it is not economical to assume that one means of encoding is more basic. As pointed out in Bouchard (2002), this view actually requires two extra sets of theoretical tools: tools to translate each secondary means into the primary means – such as scrambling features and a displacement operation – and tools to prevent the computational system from directly accessing Case markings for interpretation, although it can access them for translation into structural positions.

The second strategy to account for this kind of functional covariation offers a more promising basis of explanation. As indicated in Bouchard (1996, 2001, 2002, pp. 382–389), temporal linearity and Case marking are two of the four modes of expression of a relation allowed by the perceptual substance of oral languages (the other two being intonation and marking of the head rather than the dependent, as in polysynthetic languages). These modes are functionally equivalent: there is no reason in their nature or in logic to assume that one is more basic than the others. The options do not constitute a list of disconnected coding modes that happen to be used by language nor are their effects the result of imperfections such as displacement. The choices are not fortuitous, but are determined by the perceptual substance. Functional covariation between Case marking and rigid order is not an aberration that requires costly additions to the theory. On the contrary, its absence would be an aberration, given the initial conditions arising from the perceptual substance: we would have to explain why no language makes use of an option that the perceptual substance provides for the language faculty.

If Case marking is linked to the nature of the perceptual substance in oral languages, as assumed under this view, a close look at the perceptual substance of sign languages should explain why such marking is not required in these languages. The visual-gestural substance involves not only the physical dimension of time, but also the three dimensions of space. An effect of this difference in dimensionality is that the visual-gestural channel has a greater potential to code information by means of a physical relation. In addition to the juxtaposition found in oral languages, in which the temporal edges of two elements are in contact, “information may be coded by physically relating two elements in space, by having them share a spatial edge, as when a directional verb shares one of its edges with a sign in locus A and is directed towards another sign in locus B with which it shares a second edge” [Bouchard (1996, p. 111)]. Note that, since a locus has a certain permanency, these spatial edges need not coincide with temporal edges: a noun may have been signed and been attributed a locus earlier in the discourse, and then several signs may have been produced before the verb is signed. As we indicated earlier, the use of loci in this way is extremely frequent in sign languages, because
it avoids the need to produce a new sign each time a referent is reactualized. Because of this inherent permanency of loci due to the spatial dimension, the additional cost of paradigms can be avoided. Paradigmatic domains introduce permanency in grammar. Thus, membership in a Case-marking paradigm identifies a particular, fixed grammatical function. As in the case of number and class/gender paradigms, there is no need for this means to introduce permanency since permanency is already available through the loci: a locus can be used to directly indicate the relation that a discourse element holds with a predicate. In Saussurean terms, an associative relation of Case marking is not necessary to provide a virtual identification of a grammatical function in sign languages because the spatial locus assigned to the referent already has a certain permanency, it remains *in praesentia*\(^\text{11}\).

3.2.7. Summary

The perceptual substance of sign languages has multiple articulators and spatial dimensions which allow a signer to directly express the first three properties of (20) – participant deixis, inclusion/exclusion, and spatial deixis, whereas the perceptual substance of oral languages cannot express these properties without recourse to special forms. Number and class/gender participate in the identification of actants in oral languages, particularly to reactualize an actant; Case markings identify the actant’s role. These properties all introduce permanency through membership in a paradigm. Sign languages do not categorize elements with paradigmatic properties such as number, class/gender, or Case because the permanency of the spatial substance allows actants to be directly reactualized and their roles identified, so there is no need for paradigms to manifest permanency.

In our view, it is no accident that oral and sign languages exhibit differences in the functioning of pronouns (reactualization) such as those listed in (20). These differences are related: they all depend on certain differences in the perceptual substances of oral and sign languages.

4. Consequences for linguistic categorization and universals

Since each modality has different physiological features and since these affect certain properties that are relevant for categorization, we have to reconsider in what sense such categories are universal. If certain categorial traits are modality-dependent, then it cannot be the case that all languages have them; even to say that some of those

\(^{11}\) Whatever status grammatical functions may have with respect to universality, it is inappropriate to define them in terms of notions that derive from one particular property of a perceptual substance, to the detriment of others. Thus, it is odd to define them in terms of Case markings or of structural relations such as sisterhood and immediate dominance (which derive from temporal ordering): these are equivalent means that the perceptual substance provides to express grammatical functions.
categories are covert in some cases makes little sense. It appears more likely that our physiology and the language faculty provide a “toolbox” from which each language can choose.

A related question is what the criteria are for establishing any particular category. Traditionally, they fall into three classes: morphophonological, syntactic, and semantic. Many scholars have already noted that semantic criteria are very hard to use with any accuracy. Moreover, some languages have very little morphophonology to rely on, while others make very little use of syntactic ordering. This becomes even more critical when different modalities are considered, since the means of expressing syntactic combinations are determined by physiological properties. What emerges from the study of sign languages, and in particular LSQ, is that what is stored in the mental lexicon is not categorized forms selected according to their category, but rather noncategorized forms. Categorization then expresses the particular function the item has in a particular context of use. Thus, numerous signs have the same formal features (configuration, orientation, place of articulation and contour of movement) whether they function as a noun or as a verb because their behavior in space indicates which purpose they serve in the sentence. Similarly, the distinction between the definite determiner and pronoun functions of a pointer is simply a difference between the initial assignment of a locus and its subsequent reuse.

Each mode of coding information in the sensorimotor apparatus is subject to extralinguistic constraints imposed by human motor, perceptual, and cognitive limitations. By assuming a grammatical system that narrowly reflects the properties of one particular perceptual substance, one is imposing on the whole of grammar limitations that are due to extralinguistic properties of that substance. For instance, the temporal linearity of oral languages is often presented as being more “basic,” and other properties of that substance such as morphological marking or intonation are “translated” into phrasal structural notions deriving from temporal linearity. This can create confusion between linguistic mechanisms that are innately determined and contingent properties of the interfacing systems. The risk of confusion increases if the “translation” takes place between languages with different modalities like oral and sign languages.

We are not saying that sign languages are very different from oral languages. On the contrary, we are saying that, despite the huge perceptual differences, if we clearly understand the import of each particular substance, we can see how both types of languages are equally determined by properties of their respective perceptual substances. We believe that this is where the similarities lie, not in “translating” properties of sign languages into properties of oral languages – an effort that is of limited value when these properties arise from the perceptual substance.

In sum, oral and sign languages are actually very similar in the fundamental principles of their syntax, but important physico-perceptual differences between their modalities determine the surface realizations of these principles in ways that make them appear very different.
References


Saussure, F. (1916), Cours de linguistique générale (Payot, Paris).


Travis (this volume)
Chapter 17

SYNTACTIC CATEGORIES IN SIGNED VERSUS SPOKEN LANGUAGES

DIANE LILLO-MARTIN

University of Connecticut

Contents

Abstract 402
1. Introduction 403
2. Lexical categories 403
3. Grammatical structures 405
   3.1. Subordination 405
   3.2. “Spatial syntax” 406
      3.2.1. Pronouns 406
      3.2.2. Verb agreement 408
      3.2.3. Analyses of spatial syntax 409
      3.2.4. Complications 411
      3.2.5. The use of space to represent space 412
   3.3. Word order 413
4. Conclusion 417
Acknowledgments 418
Appendix. Notational conventions 418
References 419
Abstract

Are syntactic categories a fundamental property of language – or of spoken language? Natural sign languages share the fundamental properties of language. Are there any significant effects of the modality of transmission – the fact that spoken languages are primarily sequential, while sign languages employ a significant spatial component?

Linguistic categories include both language-specific and language-universal types. By considering the properties that languages share, linguists develop theories which attempt to capture the permissible range of variation. Most of these theories are based on the analysis of spoken languages only. Thus, they are in danger of overlooking the properties exhibited by language in the visual modality. This chapter focuses on morphosyntactic properties which on first examination seem to show effects of the linguistic modality. On further analysis, these properties are argued to be more deeply similar across the modes.
1. Introduction

What are the “categories” of syntax? Are the categories of syntax constant across languages – or across language modalities? Is there something special about language underlying its categories? These questions are among those that interest cognitive scientists because of the central role language currently plays in our understanding of human cognition. A good deal is known about the character of human language, and its investigation is a central component of current cognitive science. This is not to say that cognitive scientists or even linguists are unanimous in their characterization of language and the role it plays in the human mind. But certain facts about language must be accounted for within any theory. This chapter is about one set of such facts.

The “categories” of this chapter are the linguistic units used in the lexicon, morphology, and syntax. They comprise elements which behave as a group, by exhibiting similar distribution, use, and restrictions. They include what are traditionally considered grammatical categories – i.e., nouns, verbs, adjectives, etc. They also include morphological units (such as agreement markers) and syntactic constituents (such as noun phrases, verb phrases, and sentences). These categories display widespread similarities across languages. Thus, while not all languages have exactly the same set of grammatical categories, languages do have words that behave as one group versus another, and across languages there are many similarities in the behavior of the noun-like group versus the verb-like group. These categories (broadly understood) are a fundamental property of languages.

The study of natural sign languages allows research to reopen the conclusion of the previous paragraph. Are the categories discussed so far a fundamental property of language – or of spoken language? Natural sign languages share the fundamental properties of language (in a way that computer “languages,” or even “body language,” for example, do not). But are they exactly like spoken languages? Or are there any significant effects of the modality of transmission – the fact that spoken languages are primarily sequential, using the vocal/auditory channel, while sign languages employ a significant spatial component, using the manual (corporeal)/visual channel?

This is the central question of the present chapter. How similar – or different – are the categories of spoken and signed languages? Where they are similar – why? Where they are different – why again?

2. Lexical categories

The first linguistic analysis of American Sign Language (ASL) appeared in Stokoe (1960). Since that time, a number of sign languages have been studied, but much research is still needed. This chapter will focus on data from ASL, but other sign languages will also be discussed and issues related to generalization will be clear.

In some of the early work on ASL, it appeared that there were, in fact, significant differences between signed and spoken languages in the nature of the categories employed.
For example, Stokoe, Casterline and Croneberg (1965) characterized many signs as both nouns and verbs: AIRPLANE and TO-FLY, SCISSORS and CUT-WITH-SCISSORS, IRON and TO-IRON. While zero conversions are found in some languages (witness the English word iron), it seemed that there were no systematic differences between nouns and verbs in this language. Furthermore, it seemed that nouns were not marked for number, gender, definiteness, or other features commonly found in the better-studied spoken languages. Similarly, verbs were not marked for tense, and differentiating between verbs, adjectives, and prepositions was muddy.

Early research cleared up the confusion between nouns and verbs. Supalla and Newport (1978) showed that a derivational relationship exists between many pairs of related nouns and verbs, such as the ones listed earlier. The forms are generally distinguished by their movement – verbs have a single, unrestrained movement, while nouns have repeated, restrained movement. This is only one example in which a distinction made in ASL proved too subtle for the earliest researchers. Not all grammatical markers are represented by independent lexical items. While spoken languages do not use movement to indicate grammatical categories, they may use some phonological pattern (for example, stress or an affix). Sign language researchers had yet to be sensitized to all linguistically relevant information in the signal.

More recently, other researchers have proposed that noun phrases are marked for number and specificity/definiteness [MacLaughlin (1997), Zimmer and Patschke (1990)]. According to MacLaughlin, definiteness is expressed through a pre-nominal determiner (glossed IX for “index”), and number is marked through the use of movement modifications of this determiner. These effects are illustrated in the examples in (1) (see the Appendix for a list of notational conventions).

\[
\text{ASL [MacLaughlin (1997)]}
\]

\[
(1) \quad \begin{align*}
\text{a. } & \quad \text{JOHN LOVE [ IX₁ WOMAN ] DP} \\
& \quad \text{‘John loves the/that woman.’} \\
\text{b. } & \quad \text{[IX₉ pl-arc ; MAN IX₁ ;“over there” ] DP KNOW PRESIDENT} \\
& \quad \text{‘Those men over there know the president.’}
\end{align*}
\]

Questions about verbs and predicates more generally were similarly cleared up. It is now commonly agreed that verbs in ASL are not marked for tense, but predicates may be marked for aspect [Klima and Bellugi (1979)], thus making ASL more like Chinese than English, but still well within the range of variation observed in spoken languages. And the fact that adjectives may serve as predicates is not at all unusual in comparison with languages like Arabic, for example.

What these examples show is that the basic grammatical categories of a sign language are not fundamentally distinct from those found across spoken languages. While some variability across languages exists, sign languages fall within the range of variability already found for spoken languages.
3. Grammatical structures

If the lexical categories of sign languages are not outside the range of variation found in spoken languages, what about more complex grammatical structures? In some domains, early limitations on our understanding of sign languages, which made them look very different, were fairly quickly disproven. In others, however, debate has continued over the years, with some researchers arguing that sign languages and spoken languages are significantly different due to modality effects. I will quickly go over an example of the first type, and then discuss two examples of the second type in more detail.

3.1. Subordination

One candidate difference discussed in early research on ASL was the claim by Thompson (1977) that ASL has no subordination. One of the ubiquitous properties of human language is recursion – the ability to embed a category of one type within a category of the same type. Subordination is a prime example of recursion. It is because English can have phrase markers like that in (2) that it can also have sentences like that in (3).

(3) John thinks that Mary believes that Fred said that Jill wants …

If sign languages had no recursion, they would be missing an important property of human language, one which is part of the reason for language’s unlimited ability to generate novel utterances. So Thompson’s claim was important for our understanding of the fundamental properties of human language.

Thompson based his claim on two points. First, he failed to find an overt marker of subordination, parallel to English that. Second, he found that ASL translations of English sentences with subordination employed conjoined rather than embedded clauses.

Liddell (1980) took up the challenge implied in Thompson’s proposal and showed that ASL does have subordination, including embedded sentential complements and relative clauses. Even though some translations of some English sentences failed to show subordination, Liddell provided examples like those in (4) to support his claim that embedding exists. He also showed that there is a marker that systematically indicates
subordination in relative clauses. This marker, however, is not a separate lexical item, but a specific facial expression, known as a nonmanual marker, glossed “r” in the example. Again, researchers had overlooked an important part of the sign language signal. Since Liddell’s work appeared, researchers have considered nonmanual markers to be part of the grammar of ASL, and recent proposals have shown that these markers are parallel to grammatical uses of intonation [Sandler (1999), Sandler and Lillo-Martin (in press)].

ASL [Liddell (1980)]

(4) DOG CHASE CAT COME HOME
‘The dog that chased the cat came home.’

In the domains discussed so far, sign languages were claimed to be missing something that spoken languages have (certain contrasts or certain types of structures). When one observes sign language, however, one might be more struck by what it has that spoken language does not: an ability to make use of different positions in space to convey meaning. Spoken languages are primarily sequential, while sign languages appear to be both sequential and spatial. The way that space is used to convey meaning has been discussed extensively in the sign language literature. For some researchers, the use of “spatial syntax” sets sign languages apart from spoken languages.

3.2. “Spatial syntax”

The area of space in front of the signer’s body is used in two distinct ways in sign languages. First, many ordinary signs are articulated in “neutral space” – thus, spatial locations are sublexical units. This aspect of space is not relevant to the following discussion. However, space is also used meaningfully in sign languages: to represent space, or more abstractly, to represent referents. I will start with the latter. The use of space to represent referents is relevant to various parts of the grammar; in particular, it is relevant to the pronominal system and the system known as verb agreement. I will first describe the use of space with respect to pronouns and verb agreement, and then turn to how this use of space can be analyzed. Then, I will bring up some complexities not discussed earlier, introducing some of the most current issues. Following this discussion, I will briefly discuss the use of space to represent space.

3.2.1. Pronouns

To refer to someone who is present during a signed conversation (including the signer), the signer can point in the direction of the referent. In ASL, the point uses an extended index finger with the palm down for personal nonpossessive pronouns; other handshapes and orientations represent possessive, honorific, and casual reference. This pointing sign, illustrated in Figure 1, acts as a pronoun that picks out the intended referent within the signed discourse. When an intended referent is not physically present, the signe establishes an
abstract location to represent the intended referent, and pointing to that location (often called a “locus” in sign language literature), as shown in Figure 2, picks out that referent. Points are equivalent to pronouns in that they pick out referents without using a name or referential expression. However, they appear to be unlike spoken language pronouns in that they cannot be used ambiguously, but must pick out a specific referent. For example, if the signer is talking about Barbara and Linda, each woman can be associated with a distinct locus in the signing space – typically one on the right side, and one on the left. If Linda is associated with the locus on the signer’s right, then pointing to the location on the right picks out Linda. There is no form of the sign which could equally pick out either woman, as with the English word *she*.

Within a particular discourse, a spoken language pronoun is usually used to identify a particular referent, but the pronoun (e.g., *she*) is the same regardless of who it picks out, for the particular category which that pronoun refers to (e.g., third-person singular females). In many situations, such as that in (5), a particular pronoun may be completely ambiguous (at least to the listener). Unlike *she*, the sign language point in (5) (variously glossed in the literature as PT, for “point,” IX, for “index,” PRO(NOUN), etc.) is not even potentially ambiguous. Sign language pronouns do not denote a class of potential referents, such as third-person singular female referents, but only the referent associated with the locus they are directed toward.

(5) Situation:
Speaker and addressee have been discussing Linda and Barbara.

a. She’s in trouble.
   *She* may pick out either woman.

b. IX TROUBLE
   IX must be directed toward the locus for one of the two women.

Sign languages are therefore distinct from spoken languages in the following important way: sign language pronouns indicate a particular referent, not a class of referents.
In addition, the number of distinct referents which can be picked out this way is indeterminate. Just as the space can be divided into an infinite number of distinct regions, so too is there potentially no limit to the number of distinct pronouns which can be articulated. In practice, no more than three or four referents will usually be distinguished using spatial loci in any one discourse segment, but this limitation should arguably be attributed to memory and perceptual limitations rather than grammar.

It is important to recognize that this phenomenon is not the same as that found in a language (such as Mixtec) that has a large number of pronouns, each of which picks out a different gender class. The pronouns in sign languages apparently identify individual referents, not classes (even very small ones). They may also convey additional information about the referent, including gender, number, and person, to be discussed shortly.

3.2.2. Verb agreement

Verb agreement in sign languages makes use of the same locations in space as pronouns. Typically, a verb marked for agreement begins at the location of its subject, and moves toward the location of its object. Agreeing verbs also generally “face” their object – that is, the forearm (radio-ulnar part) twists so that the palm and/or fingertips are oriented toward the locus of the object. Furthermore, the signer’s head may turn so that the eyes gaze toward the locus of the object [Bahan (1996)]. Verb agreement is illustrated in Figure 3, with two forms of the sign ASK. If the signer herself were associated with the locus on her left side Figure 3a would be interpreted as ‘I ask him.’ If George were associated with the locus on the signer’s right, Figure 3b would be interpreted as ‘He asks her.’ As with pronouns, the referents may be present or nonpresent, but the verbs are modified in the same way in either case.
3.2.3. Analyses of spatial syntax

In what sense is the system just described one of agreement? Since noun or pronoun signs may be signed in or directed toward loci, and verbs are also signed with respect to these loci, verbs “agree” in this way with their arguments. Furthermore, verbs marked with agreement behave in various ways like agreeing verbs in spoken languages; for example, they license null arguments [Lillo-Martin (1986), Bahan et al. (2000)]. But what features of the subject or object nouns are marked on the verb? First consider the common agreement features of gender, number, and person.

In ASL, neither nouns nor verbs are marked for gender. However, it seems that some sign languages do have gender marking. For example, Taiwanese and Japanese Sign Languages are said to mark gender on certain verbs, by changing the handshape of the nondominant hand [Fischer (2000), Smith (1990)]. Clearly, even for sign languages which do mark gender, this does not explain the spatial system described previously.

Sign languages also convey number on pronouns and verbs. A class of plural referents can be depicted in two ways. One way is to associate a group with one locus; hence, for example, the students in a class can be associated with one locus, and a point or movement of an agreeing verb toward that locus picks out the students as a whole. In this case, the pronoun or verb picks out a group, but no information about number is marked on the pronoun or verb itself. This is not an example of grammatical number.

Alternatively, plural referents can be identified by the addition of an arc movement to a pronoun or verb, illustrated on the verb ASK in Figure 4. (There are also separate exhaustive and dual forms, not discussed here.) The multiple form can only mark the verb’s object, and there are further limitations on possible combinations of particular verbs, locations, and numbers. These limitations can largely be attributed to morphological and/or phonological constraints [Mathur (2000)]. Thus, while “arc” can be considered the representation of the feature ‘plural,’ number is apparently not responsible for the spatial aspects of verb agreement either.
What about person? Are the various forms used to pick out different referents equivalent to differences in person? This has been the prevailing view, in some form or other, in the sign language literature. For example, Padden (1988) considered verbs as agreeing in person and number with the subject and object; she assumed that person distinctions include first, second, and third. Meier (1990) argued that ASL only makes a distinction between first and nonfirst. While the first-person referent has a specifiable location (the signer), which shifts its reference in direct discourse, and idiosyncratic forms for first plural, there are no linguistic reasons to separate second from third person. Thus, Meier concludes that ASL has two person categories: first and nonfirst.

Of course, the claim that there are only two person categories does not seem to account for the multiplicity of forms actually found. A pronoun or agreeing verb may be directed toward any location in space. How are all the various nonfirst forms to be distinguished?

Neidle et al. (2000) make their position on the matter clear. They say, “We claim that such use of spatial locations constitutes an overt instantiation of φ-features (specifically, person features) associated with these referential entities, since these locations in space systematically participate in the same linguistic phenomena that involve φ-features crosslinguistically” [Neidle et al. (2000)]. And in a nearby footnote, they clarify: “Here we propose that, although (consistent with Meier’s claim) there is a primary distinction between first and nonfirst persons, nonfirst person can be further subclassified into many distinct person values” (pp. 166–167). But how many distinct person values can nonfirst be subclassified into? The number would seem to be nonfinite.

Lillo-Martin and Klima (1990), recognizing the difficulty of such a position, make a somewhat different proposal. If each locus corresponds to a distinct person value, this would require the postulation of an indeterminate – or unlimited – number of person distinctions in the lexicon, something they find unacceptable. In this paper, which only discusses the issue with respect to pronouns, they propose that the distinct pronominal forms found on the surface in sign languages reflect a single pronoun with no person marking, but with an overt realization of the abstract referential index that they assume all pronouns bear.

Suppose, following the Binding Theory view [Chomsky (1981)], that every NP (or DP) is marked with a referential index. This index is used to check compliance with the binding
theory, and it is relevant to interpretation. In spoken languages, it may play a role in pronunciation, as when a pronoun that is coreferential with an NP in its sentence is stressed, as in example (6). Generally, however, there is no overt instantiation of this index.

(6) GEORGE, thinks HE, is the best candidate.

Lillo-Martin and Klima suggest that this referential index is overt in sign languages so that NPs (pronouns) with the same referential index are associated with the same locus. In this view, it is not necessary to say that sign languages are qualitatively different from spoken languages with respect to their pronominal systems. The only difference is in the pronunciation – where differences are to be expected.

For maximum generality, Lillo-Martin and Klima suggest that sign languages make no person distinctions at all in their pronominal system. However, in subsequent work, I have adopted an analysis which attributes to sign languages a first/nonfirst distinction (following Meier), together with overt realization of referential indices.

An extension of this view to the verb agreement system would have the $\phi$-features with which a verb agrees consist of first vs. nonfirst person, number, and (in certain sign languages) gender – but not the distinction between individual referents indicated by distinct spatial loci. This distinction would be represented in the syntactic analysis by a referential index rather than by $\phi$-features.

3.2.4. Complications

There are some remaining problems with the description given so far. It is an oversimplification in many ways. One problem is that not all verbs participate in the agreement system. Padden (1988) categorized verbs into three classes: “agreeing” (originally called inflecting, but now called agreeing or agreement by most researchers), “spatial,” and “plain” (nonagreeing). According to Padden’s analysis, the classification of verbs is largely a matter of lexical specification. Agreeing verbs participate in the agreement system just described. Some agreeing verbs are “backwards,” because they move from the location of the object to the location of the subject. Spatial verbs, such as GO-TO, PUT, and MOVE, have locative arguments, and their movements indicate source and goal. Plain verbs do not indicate their arguments using spatial modifications. (This too is an oversimplification, but it is sufficient for the present purposes.)

Attempts to make these issues less ad hoc were made by Janis (1992, 1995) and Meir (1998, 2002). Meir’s analysis extends the furthest, accounting for both the classification of verbs and the patterning of backwards verbs by reference to the verbs’ lexical conceptual and predicate argument structures. First, she observed that agreeing verbs all involve literal or metaphorical transfer from a source to a goal, typically subject to object. The backwards verbs are backwards because they involve transfer from the object to the subject. Meir also observed that while the path movement of backwards verbs is, indeed, “backwards,” their facing (the orientation of the palm and/or fingertips), like that of regular verbs, is toward the object.

According to her analysis, both agreeing verbs and spatial verbs have in their Lexical Conceptual Structure (LCS) a directional morpheme, glossed DIR, which denotes the
path or trajectory a referent traverses. It is the DIR morpheme which takes agreement, moving from source to goal. In addition, agreeing verbs (including GIVE, ASK, and HELP) are categorized as verbs of transfer, such that the theme (element being transferred) moves along this path from a source to a goal.

Using this analysis, several facts are accounted for. One is that sign languages seem to share the classification of verbs, which does indeed fall along lexical/semantic lines (with exceptions which are phonologically based). Interestingly, Meir’s analysis is based on data from Israeli Sign Language – but it seems to apply perfectly to ASL as well. Another is that backwards verbs (such as TAKE, COPY, and STEAL) are exactly those whose lexical semantic structure involves transfer from the object to the subject. Noticing this allowed Meir to capture the underlying features of these verb types.

While recent work has shown greater regularity and predictability in the system of ASL verb agreement than the Padden-type classification, some challenges still remain. Over the last 15 years, Liddell (1990, 1994, 1995, 2000) has noted the difficulty of morphologically specifying the loci associated with nominals and employed by agreeing verbs. Arguing that “[t]he concept of a lexically fixed, meaningful element with indeterminate form is inconsistent with our conception of what morphemes are” [Liddell (1995), p. 25], he concludes that there is no linguistic process of verb agreement at all in ASL. Instead, he argues that the modification of verbs to “indicate” referents comes from the gestural component, intricately combined with the linguistic elements of the signs.

Liddell has a point. The problems are even more complicated than I have presented them here. However, there are good reasons to consider the modification of verbs in ASL as constituting agreement. Responses to Liddell’s arguments have been presented by Aronoff, Meir and Sandler (2000), Lillo-Martin (2002), Meier (2002), and Rathmann and Mathur (2002). In general, these responses have argued that linguistic (and psycholinguistic) evidence supports the notion that agreement (for first vs. nonfirst person, and number) exists in sign languages. Some have, however, accepted Liddell’s argument that a full description of the form of agreeing verbs must take into account both linguistic and gestural components.

### 3.2.5. The use of space to represent space

Liddell’s argument goes beyond the system of verb agreement in sign language, because space is “syntactic” in sign languages in other ways. In particular, the signing space can be used to represent not only referents, but also physical spatial relations. Verbs such as COME and GO (which Padden classified as “spatial” verbs); “classifier” verbs, which combine information about the size and shape or semantic class of their arguments with movement and/or location predicates; and even “plain” verbs may use spatial locations to represent spatial locations. There have been fewer attempts to come up with a syntax of the use of space in such verbs, although they have received considerable attention [see, e.g., Benedicto and Brentari (1986), Emmorey (1990), Schick (2003), Supalla (1986)].

These topics are of theoretical interest, because they clearly must be considered as potential evidence for modality effects. Their ubiquitous presence across sign languages
shows that they are effects of the visual modality – and thus, require explanation. Careful attention to these issues will allow researchers to refine the line between linguistic and gestural, and to identify true linguistic universals.

3.3. Word order

During the 1970s, researchers questioned not only whether ASL syntax showed recursion, but also whether or not it was possible to define a basic word order for the language. Several researchers had observed that word order in ASL is fairly flexible. Fischer (1975) claimed that the basic word order of ASL was Subject-Verb-Object (SVO), but she reported that alternative orders are possible in various circumstances, including semantically “nonreversible” sentences, such as (7); and sentences with verbs marked for agreement with subject and object, such as (8). She also observed that orders other than SVO might be used even when the nouns are semantically reversible, but that in such cases there would be an intonation break indicating that the sentence-initial information was topicalized, as shown in (9).

\[
\begin{align*}
(7) & \quad \text{a. BOY LIKE ICE-CREAM} \\
& \quad \text{b. BOY ICE-CREAM LIKE} \\
& \quad \text{c. ICE-CREAM LIKE BOY} \\
& \quad \text{‘The boy likes ice-cream.’}
\end{align*}
\]

\[
\begin{align*}
(8) & \quad \text{a. GIRL KICK BOY} \\
& \quad \text{b. BOY (HERE) GIRL (HERE) SHE-KICK-HIM} \\
& \quad \text{c. GIRL (HERE) BOY (HERE) SHE-KICK-HIM} \\
& \quad \text{‘The girl kicked the boy.’}
\end{align*}
\]

\[
\begin{align*}
(9) & \quad \text{a. MAN NOTICE CHILD} \quad (\text{SVO}) \\
& \quad \text{‘The man noticed the child.’} \\
& \quad \text{b. CHILD, MAN NOTICE} \quad (\text{OSV}) \\
& \quad \text{‘As for the child, the man noticed it.’} \\
& \quad \text{c. NOTICE CHILD, MAN} \quad (\text{VOS}) \\
& \quad \text{‘He noticed the child, the man did.’}
\end{align*}
\]

Friedman (1976) argued against Fischer’s analysis, claiming that word order is simply relatively free, based on her observation that SVO is relatively infrequent in the discourses she analyzed. However, Liddell (1980) and Padden (1988) argued against this view, showing that basic SVO word order, plus movement operations including topicalization, better accounted for a range of facts, including word order in yes/no questions, and sentential complements. Eventually, the field settled on the conclusion that the underlying order of ASL is SVO.

However, the notion that (some) sign languages may rely on discourse factors rather than sentence-level syntax for word order has been supported by some researchers [see,
e.g., the papers in Brennan and Turner (1994)). According to this view, however, sign languages might not be categorically distinct from spoken languages, but parallel to those discourse-oriented spoken languages such as Chinese or Russian that also make use of discourse or information structure in determining order. Different theories would make use of different theoretical mechanisms to account for the surface patterns in such languages – some using discourse mechanisms to directly generate the surface orders, and others using movement rules such as topicalization and focus-movement. In either case, the claim would not be that sign languages are systematically distinct from spoken languages, but that they display a type of structure used in some (if not all) spoken languages.

A different view was put forth by Bouchard and Dubuisson (1995). Reviewing the literature on basic word order in ASL, they noted that many authors recognized more flexible word order in ASL than is found in a language like English. While, as we have seen, various proposals have been made to account for the variations in word order found, Bouchard and Dubuisson concluded that order is just “one means among others of expressing the combination of elements” (p. 108). When other means – such as verb agreement – are used, according to Bouchard and Dubuisson, word order is free. In particular, they argue against an account that relies on an underlying order plus order-changing operations to derive different orders; rather, they claim, in such circumstances “order is free, since it is not functionally necessary” (p. 109).

A problem for this conclusion is the fact that order is not completely free, even with agreement or spatial information. VSO is not a permitted order in ASL, for example. And the extent to which non-SVO orders are permitted is actually a matter of some dispute. While sequences of SOV or OSV such as those in (8) are sometimes found with verbs showing agreement, they are not completely free. It has even been argued that such sequences actually constitute three distinct clauses: Noun-1 is here; Noun-2 is here; null-subject Verb null-object [Padden (1988)].

There are interesting cases in which SOV or OSV may be found with verbs marked with spatial, aspectual, or classifier information. But these cases may be analyzed in relation to another order found with such constructions, SVO [cf. Fischer and Janis (1990), Matsuoka (1997), Braze (2004)]. Such analyses start with the basic, underlying SVO order and provide some mechanism for the doubling of the verb, or the raising of the object, given a particular morphological context. There is variability in the acceptance of some examples, but examples like (10) are good for many signers.

(10) a. STUDENT NAME S-A-L-L-Y TYPE HER TERM PAPER TYPE[cont]
   ‘A student whose name is Sally was typing and typing her term paper.’

b. SALLY PAPER TYPE[cont]

c. PAPER SALLY TYPE[cont]
   ‘Sally was typing and typing her paper.’

There is also plenty of evidence that, with or without special morphology on the verb, in many cases non-SVO orders are associated with particular discourse markers.
indicating topic (11) or focus (12) for example [see, among many others, Liddell (1980), Fischer (1990), Aarons (1994), Wilbur (1994, 1996, 1997)]. While the constituents in noncanonical position may be analyzed as moved or base-generated, this type of behavior is far from making signed languages unique. Rather, they seem to behave like those spoken languages which use discourse structure or morphological information to permit variation from the underlying order. In ASL, these alternative orders are associated with both a particular discourse context, and a prosodic cue – and they show other classical effects of derived structures (such as constituency tests and constraints on long-distance relations) [Padden (1988), Lillo-Martin (1991), Petronio (1993)].

ASL [Liddell (1980)]

(11) Topic

\[
\begin{array}{l}
\text{CAT, DOG CHASE} \\
\end{array} \\
\text{‘As for the cat, the dog chased it.’}
\]

ASL [Wilbur (1996)]

(12) Focus (wh-cleft)

\[
\begin{array}{l}
\text{KIM SEE STEAL TTY WHO, LEE} \\
\end{array} \\
\text{‘Kim saw that the one who stole the TTY was Lee.’}
\]

Bouchard and Dubuisson’s remark, however, was not intended solely to characterize signed languages as distinct from spoken languages with respect to order. Instead, they advocate a much stronger position [expanded on and clarified in Bouchard (1996)], which is that languages may vary in the extent to which they rely on structure to provide information about the relationships between elements in a sentence (or discourse). This is not simply a claim that sign languages employ the same hierarchical structure as spoken languages in the syntax, but permit greater variability in ordering elements. Instead, they advocate the approach that all languages may choose from two universally available means for indicating the establishment of a relationship between elements. The first means is to put the elements in a physical relation, so that they share an edge or space. For spoken languages, this can only be done by temporal ordering. For sign languages, however, this can be done by temporal sequencing or by using the same physical space for the elements. Signed languages are distinct from spoken languages in that they may (and invariably do) make use of this latter mechanism. The second mechanism is modification of the elements, such as is found in Case-marking. When means such as Case-marking or physical locations are taken advantage of, Bouchard claims, structure is unnecessary – and so the most minimal conception of the human language mechanism would omit structure in these cases.

Bouchard and Dubuisson argue that structure-based analyses of ASL have not proven the necessity for structure rather than some other mechanism such as the one
they propose. In particular, they argue against the proposals by Aarons et al. (1992), which include a complex phrase marker containing multiple functional categories with mixed headedness. The proposals by Aarons et al. (as well as other works of theirs), defended in Kegl et al. (1996), were not intended to argue for the notion that sign language sentences have structure; rather, presupposing this common idea, they apply a particular structural proposal to ASL data. While I dispute many of their specific hypotheses [see Sandler and Lillo-Martin (in press)], I agree with them that completely dispensing with the notion of structure is unwarranted, and fails to account for the full range of data.

For example, even Bouchard and Dubuisson (1995) see that structure and order are required for sign language sentences that do not use spatial relations, such as utterances with plain verbs. They “conclude that ASL is a mixed language, with an ‘Elsewhere’ approach to order” (p. 109). That is, if the use of spatial locations allows all relevant information to be expressed without order, then order is free, but elsewhere ASL reverts to order. This conclusion puts Bouchard and Dubuisson in a rather awkward position. If sign languages do not need order because they have space, why do they sometimes need order after all? And what is it that determines the “elsewhere” order? And why should sign languages, like spoken languages, vary with respect to what looks like the “elsewhere” order (ASL is SVO, but German Sign Language – Deutsche Gebärdensprache (DGS) – is SOV)? Bouchard and Dubuisson’s argument is that if order is not needed, a minimal conception of the human language capacity would not include it. But it is not the case that order is not needed; since it must be part of the human language capacity, there is not necessarily an additional cost to the assumption that it is used in all cases. Bouchard and Dubuisson are right to ask whether the analyses which assume structure make use of too many additional, unwarranted, assumptions in order to derive the correct structure. But arguing against particular analyses within a generative framework that includes structure is different from arguing against the whole idea of structure.

This debate reminds me of the state of affairs in the formal linguistic investigation of various so-called “free word order languages.” One example comes from the study of Japanese in the early 1980s. Some authors claimed that Japanese is “nonconfigurational” [see Hale (1980, 1983), among others], because of its apparently “free” word order and other properties. A broad set of facts about the language could be captured using this notion. However, Hoji (1985) and Saito (1985), among others, showed that hierarchical structure is required in Japanese, and employed the movement operation known as scrambling to account for the variations in word order found (the apparently free operation of scrambling allows elements to be reordered in Japanese to a much greater extent than that found in ASL – but still, even in scrambled Japanese sentences there are ordering restrictions, as the verb must be final).

One source of support for the claim that hierarchical structure is a crucial component of Japanese comes from binding asymmetries, which establish that Japanese has a VP setting up an asymmetry between the subject and scrambled elements, which are higher in the structure, and the object, inside VP. For Saito, an example like this established
the existence of VP at S-structure, and by the Projection Principle, at all levels of structure. These days, it is often assumed that the condition relevant to this structure applies in Logical Form, and no Projection Principle is assumed, so this type of example alone would not imply the existence of VP within the syntax. However, Saito gave additional evidence, from crossover and quantifier float, for the existence of VP in Japanese and the application of scrambling to derive orders other than SOV, and so much evidence has accumulated since then that Japanese is no longer considered a candidate for “non-configurationality.” Saito quotes from Kuroda:

Word order is quite free in Japanese. This is obvious. But to transform this trivial observation of a phenomenon into a grammatical principle is another matter. There are good indications that the linear order of certain constituents is grammatically relevant, if one pays attention to the interaction between so-called word order and certain other grammatical phenomenon.


The debate about configurationality in Japanese has been repeated for numerous other languages, and the conclusion has always been in favor of hierarchical structure. I believe the same thing will be true of ASL (and probably other sign languages). Arguments for hierarchical structure and movement rules to derive non-SVO order go back at least to Fischer (1974) and Liddell (1980). Certainly more work needs to be done in this area, directed at two issues. First, future research may provide noncircular, conclusive, “knock-down” evidence—like that which Saito and Hoji provided for Japanese—that hierarchical structure is required to account for sign languages, even in situations with spatial information. Second, future researchers may ask whether sign languages can offer any new insights into the recent idea that while hierarchical structure is part of syntax, ordering is not. What I have argued is that sign languages do not differ from spoken languages either in their need for hierarchical structure in syntax, or in the existence of linearization at some level. Can sign languages provide any new information regarding whether that level is Phonetic Form?

Whatever the answer to the second question may be, the discussion in this section supports the claim that structure dependence is a criterial component of language, whether spoken or signed. If this is borne out, the case for structure in the human language faculty becomes all the more compelling. Bouchard’s hypothesis emphasizes that, if it were possible for the human language computation system to work without structure, this would be apparent in sign languages. If even these languages require hierarchical structure, it must be a primitive.

4. Conclusion

Linguistic categories include both language-specific and language-universal types. By considering the properties that languages share, linguists develop theories that attempt to capture the permissible range of variation. Most of these theories are based on the analysis of spoken languages only. Thus, they are in danger of overlooking properties
exhibited by language in the visual modality. A putative linguistic universal can be valid-
ated or contradicted by examination of the properties of sign languages.

This chapter has focused on morphosyntactic properties which at first glance seemed
to show effects of the linguistic modality, but which were argued to be more deeply sim-
ilar underlyingly. However, some modality effects were already apparent, such as the
overt realization of indices on pronouns and agreeing verbs. Areas of potentially more
profound difference were alluded to, including the use of spatial information to convey
information about space.

An additional area of potential modality effects concerns the strong similarities
across sign languages. While spoken languages exhibit various typologies, sign lan-
guages also display some intriguing clusters of properties [Newport and Supalla
(2000)]. For example, as far as we know, sign languages as a whole display the character-
istics of discourse-oriented languages, while only some spoken languages are so
classified. It is possible that future research will show that sign languages exhibit vari-
ability in the same way that spoken languages do, but it also seems possible that there
will be a need to explain why sign languages behave as a group in certain ways [see
Sandler and Lillo-Martin (in press)].

Linguists have debated – and will continue to debate – many aspects of the structure
of sign languages and how to analyze them. It is clear, however, that sign languages
must be a part of the database on which theories of language, and the mind, are built.

Acknowledgments

The preparation of this chapter was supported in part by National Institutes of Health
Grant #NIDCD DC-00183. Portions of it appeared previously in Glot International
(Volume 5, 2001).

Appendix. Notational conventions

The following conventions are used for sign language notation. One of the concerns of
current research is that glosses fail to adequately represent all relevant information.
However, the following usages are standard.

BOY Signs are glossed in all caps.
IX Pointing signs used for pronouns.
¶ Nonmanual markers are indicated by a solid line above the glosses for the signs
they co-occur with. “r” indicates a nonmanual marker associated with relative
clauses; “t” indicates a topic nonmanual; “br” indicates a brow raise.
SHE-KICK-HIM The use of this notation in example (8) follows the notation used in
the source for a verb with subject and object agreement.
V[+] A verb with some kind of modification, including agreement, spatial, aspectual, or classifier information, is marked [+].

V[cont] This notation in example (10) indicates a verb with continuative aspect.

References


Chapter 18

ON SYNTACTIC CATEGORIES

MARK C. BAKER

Rutgers University
The four chapters that make up this section are beautifully balanced in focus and perspective. Two of the chapters discuss syntactic categories in Austronesian languages (Gil and Travis); the other two consider categories in sign languages (Bouchard et al. and Lillo-Martin). Furthermore, two of the chapters emphasize how different category systems can be in different kinds of languages (Gil and Bouchard et al.); the other two present evidence that the same category distinctions found in English and its relatives can also be found in these more exotic languages, once one knows what to look for (Travis and Lillo-Martin). Taken together, the four chapters explore two sides of an important theoretical question, each illustrated with examples drawn from interesting and readily comparable languages. They therefore constitute a near-perfect microcosm of a broader debate.

David Gil’s chapter claims that some languages have a much simpler grammatical apparatus than the familiar Indo-European languages do, morphologically, syntactically, and semantically. He believes that artificial semiotic systems, human language at an earlier evolutionary stage, and early child language all exemplify this particularly simple type of “Isolating-Monocategorial-Associational” Language. But the case that he discusses in some depth is Riau Indonesian (RI). He shows that this language is very simple morphologically, having only three affixes, a few clitics, and limited compounding. He claims that the language has only one open-class syntactic category, which he calls “S”; any two Ss can combine to form a new S, with no significant differences in structure or distribution. Finally, the core semantics of RI is merely associational: the structure [S S1 S2] is uniformly interpreted as “something associated with S1 and S2.” Although RI also has a few fixed items and the important notion of headedness, Gil claims that its grammar really is almost this simple – and hence that natural languages can vary radically along these dimensions.

Lisa Travis’s chapter presents a broad overview of the issues relevant to a theory of lexical categories from a more standard generative perspective. She discusses the pros and cons of dividing up the four lexical categories into natural classes using two distinctive features, and she reviews how the intertwined theories of functional categories, selection, and head movement can explain familiar differences in how nouns and verbs are inflected. She goes on to point out the importance for an overall theory of syntactic categories of two less-studied kinds of categories: “crossover categories,” which can (for example) transform a verbal projection into a nominal projection at some level of structure, and “multifunctional categories,” in which a single lexical item can be inserted into different syntactic positions with predictably different results. She illustrates these with gerunds in English, nominalizations in Malagasy, and the tense-aspect system of Igbo. She closes her chapter with an implicit response to work like Gil’s, showing that although Malagasy (like RI, an Austronesian language) might look at first glance like a monocategorial language [see her examples (43) and (44)], a deep knowledge of the language reveals subtle differences between nouns, verbs, and adjectives. In particular, she shows that the unusual process known as Agent Raising applies in a significantly different way to lexical items that are intrinsically adjectival and those that are intrinsically verbal.
Bouchard, Dubuisson, and Parisot’s chapter (BD&P) considers syntactic categories in what is arguably an even more “exotic” language than RI: Quebec Sign Language (LSQ). Like American Sign Language (ASL) and other sign languages, LSQ is manifest in a very different medium from spoken languages, employing a set of articulators with different physical properties (the hands, arms, and face, rather than the tongue, lips, and vocal cords) and a very different perceptual system (vision, rather than hearing). Comparing sign languages to spoken languages should then help us to see what is intrinsic to the central language system versus what should be attributed to the practical realities of the channel through which it is expressed. More specifically, BD&P argue that the different sensory modality used by LSQ results in a different syntactic categorization from spoken languages, as sign languages do not distinguish noun roots from verb roots or pronouns from determiners, and do not employ subcategorial inflectional distinctions like number, gender, and agreement. They claim that sign languages do not need these grammatical devices because of the special opportunities that the visual mode offers – in particular, the opportunity to use multiple articulators simultaneously, and access to three-dimensional space that persists over time. Their chapter is somewhat similar in spirit to Gil’s inasmuch as it also wants to do justice to how different natural human languages can be.

Diane Lillo-Martin’s chapter also focuses on a sign language, ASL, to consider how the difference in modality might impact the syntactic categories in the language. But her conclusions are quite different from BD&P’s. She reports that, as knowledge of ASL has increased, the important differences claimed to exist between ASL and spoken languages have generally decreased. ASL was once thought not to distinguish nouns from verbs, not to have subordinate clauses, and not to mark number and definiteness. However, all these features have since been found in the language, as linguists have realized that things like movement and facial expression are actually part of the linguistic representation in sign languages. Lillo-Martin also discusses in some detail the issue of word order in ASL, disputing BD&P’s claim that the richness of the visual modality makes hierarchical structure and word order unnecessary. ASL, she claims, is a rather normal subject–verb–object language, and the deviations from this order that are found are no different in kind from those found in topic-prominent spoken languages like Chinese or scrambling-rich spoken languages like Russian. Her results do converge with BD&P’s in showing that there is something special about how pronouns and determiners exploit three-dimensional space in sign languages. But with this relatively circumscribed exception, Lillo-Martin concludes that even “morphosyntactic properties which on first examination seemed to show effects of linguistic modality” turn out to be “more deeply similar across the modes.”

Considering all this material together, what are we to make of the controversy about whether natural human languages can differ radically in their systems of syntactic categories? Gil and BD&P have an explanation for why this controversy exists and why it has been so difficult to resolve: they attribute this to the failure of many linguists to look seriously at the relevant languages without the biases implicit in their Eurocentrism [see especially Gil’s article]. This is a serious issue, and one that we must constantly be
aware of. But I believe that the danger of researcher bias cuts both ways. There are powerful temptations that could lead linguists to exoticize the language they are studying, playing up differences that are not really there. These can be just as powerful as the temptations to domesticate the language of study, ignoring differences that really are there. For example, exoticizers can claim to be more sensitive to and understanding of the people they are studying. They can claim that their language of study is unique, providing a dramatic new window on human nature, and therefore their work on it is worthy of additional financial support and attention from the linguistics community. And they can have the joy of calling others ethnocentric. I know all this from personal experience, because I find myself subject to both the temptations of the exoticizer and those of the domesticator (and have fallen prey to each). The best solution that I know is to identify both temptations, and try to balance them. If I convince myself that I will be equally happy if my language turns out to be radically different (because then it is new and exciting) and if it turns out to be familiar (because then it is understandable and confirms existing theories), then I can hope to take myself out of the equation and be ready to learn the truth of the matter.

Trying to set aside our biases in this way, what else can we make of the controversies about category systems? On this point, I will take sides. The linguistic descriptions that Travis and Lillo-Martin use to infer the universality of the familiar category distinctions look deeper to me than the descriptions that Gil and BD&P use to infer radical differences in categorization. For example, Gil concludes that RI is (nearly) monocategorial because one can say both “eat chicken” and “chicken eat,” with either form referring to the chicken or to the event of eating. That is a relatively simple and superficial phenomenon. In contrast, Travis concludes that Malagasy has an adjective–verb distinction, because she discovered a surprising cluster of differences when investigating the complex phenomenon of “Agent Raising,” originally discovered by Keenan and Ralalaohery (1998). Logically speaking, if a categorical distinction manifests itself anywhere in the grammar of a language, that is enough to refute monocategorialism and to establish that there is a real distinction in the language. Examples like Travis’s, although rare and complex, should thus be sufficient. Gil writes: “So far, I have found no evidence for syntactic categories other than the open category S and the closed category S/S, and no evidence for syntactic structures other than those that can be built up recursively from subtrees such as those in (9).” That is good as far as it goes, but one can legitimately wonder how hard he has looked. Has he looked hard enough to discover a difference like Agent Raising, if something comparable were to exist in RI? Similarly, Lillo-Martin discusses how some of the distinctive properties of syntactic categories that BD&P mention for LSQ were once held to be true of ASL too, but as research has progressed on the language, more and more familiar structures have been identified. Notice how the languages that are claimed to have radically different category systems are those that have been studied for about 10 years by a single team of researchers, whereas the languages that are claimed to have familiar category systems are those that have been studied for several decades by communities of researchers (including native speaker-linguists). I do not want to convey disrespect for Gil and
BD&P: I am very glad that they have the initiative to be among the first to study important new languages, and the courage to entertain radical hypotheses about them. But like Lillo-Martin, I notice a trend that the more we know about a language, the less “exotic” it seems.

This is certainly true of my own thinking about lexical categories [Baker (2003)]. I entered into my study reasonably confident that I would find important differences in the lexical category systems of the Amerindian and African languages I know best. At one point, I was ready to present arguments for these differences at a public lecture, based on my study of the relevant literature. But two days before the lecture, I was lucky enough to have extended meetings with experts and native speakers of three key languages. In every case, familiar category distinctions came to light, completely reversing my conclusions. I was a bit saddened not to have discovered interesting new parameters for linguistic theory, but I comforted myself with the thought that I had supported Universal Grammar. I believe that this was not a bias that I brought to the work, but rather an empirical result that the work discovered.

Even if it does turn out that all languages have, at some level, the familiar category distinctions between nouns, verbs, and adjectives, it remains true that these distinctions are much more subtle in some languages than in others. We can still ask what is the difference between a language like English or Latin, where no one could miss the distinction between a noun and a verb, and Austronesian languages and sign languages, where it is very easy to miss. This difference seems real, and it demands its own kind of explanation. My own work on this issue suggests that, to clarify these matters, one has to be very sensitive to the presence of functional categories in a structure (see Travis’s chapter for some background). The following paradigm from Nahuatl is, I believe, instructive:

   1sS-3sO-make the shoe-nsf
   ‘I make the shoes.’

   1sS-3sO-make the good-nsf
   ‘I make the good one.’

c. Ni-cac-chi:hua. [Launey (1981)]
   1sS-shoe-make
   ‘I make shoes.’

d. *Ni-cual-chi:hua.
   1sS-good-make
   ‘I make good things.’

Nahuatl is often considered to be a language with no adjective-noun distinction. This is justified by the similarity between (1a) and (1b), where the root meaning “shoe” and the root meaning “good” are equally appropriate when used as the direct object of a verb like “make.” But the comparison is confounded by the fact that these lexical items
do not constitute the object by themselves; rather, they combine with the noun suffix –tlí and the determiner-like particle in. These functional categories could have nominal properties of their own, concealing the inherent qualities of the roots. Examples (1c) and (1d) are evidence that this is true. Nahuatl allows productive incorporation of the direct object, and when incorporation occurs, the object is predictably stripped of all functional material [see (1c)]. In these incorporation structures, a clear noun–adjective distinction appears: the adjectival root ‘good’ cannot incorporate in this way (1d), even though the noun root “shoe” can. The contrast between (1c) and (1d) is perfectly parallel to the English distinction between I made shoes yesterday and *I made good yesterday, and it can be explained in the same terms [see Baker (2003) for details]. Like Travis, I conclude that (many) lexical items have inherent syntactic categories as part of their lexical entries.

What does this perspective predict about languages where category differences are hard to find? We expect these to be languages in which it is hard – conceivably even impossible – to isolate lexical categories from the functional categories that surround them. There could be two quite different ways this could happen. First, a language might require functional categories to be present for language-particular reasons. For example, if Nahuatl were like French in requiring determiners to appear with all nouns and in having no object incorporation, we might be unable to isolate the difference between nouns and adjectives. This is roughly the case in Salish languages. Second, a language might have many functional categories that are phonologically null, making it hard to tell by inspection whether a confounding functional category is present or not. We know that individual functional categories can be null in individual languages, so there exists the logical possibility that some languages might have quite a few such categories, making it hard to see the category differences that are there underlyingly. RI seems to be just such a language, containing very few overt functional categories, perhaps as a result of a degree of pidginization/creolization. (Gil denies that RI is a Creole, but he describes it as a language used as a lingua franca for interethnic communication in an ethnically heterogeneous region full of recent migrants, the classic circumstances for creolization.) We can thus identify a conceptual reason why a kind of “isolating” language would look like a monocategorial language.

Turning to sign languages, what should we make of the profoundly different way that pronouns work in these languages, a point on which BD&P and Lillo-Martin agree? This is a fascinating difference, and an area where the difference in modality might indeed reveal something deep about the human language faculty. It does not look like a difference in the inventory of syntactic categories; sign languages apparently do have pronouns, and the pronouns are systematically related to determiners, as they are in spoken languages (see BD&P). Rather, it seems to be a systematic difference in the internal structure of the category pronoun, and in how instances of that category are used. English and all other spoken languages have a fixed stock of third-person pronouns (he, she, it, they), and as a result what those pronouns refer to is often ambiguous. In contrast, all sign languages make up new pronouns on the spot, simply by pointing to a new location in space. The striking result, mentioned by Lillo-Martin, is
that pronouns are never ambiguous as to their reference in sign languages, potentially making the whole topic of binding theory moot in such languages.

Does this striking difference follow from the intrinsic differences between the visual and aural modalities? The answer must be yes, it seems, given the consistency in the difference between spoken languages and signed ones. But I suspect the real reason is a bit less direct than BD&P and Lillo-Martin suppose. BD&P emphasize the permanency of the visual space as a factor in this, in that a person can keep pointing to the same location. They also take the continuity of space to be relevant, in that people can point to an arbitrary number of visually distinguishable locations. But those do not seem like the crucial distinctions. The visual trace of a pointing gesture is every bit as ephemeral as the sound waves associated with uttering a word. Moreover, the range of different sounds that can be discriminated by the human ear is probably every bit as great as the range of pointing gestures that can be distinguished by the human eye. What would a sign-language-like pronoun system look like, when translated into a spoken modality? I can imagine a language which used variants of the natural numbers (one, two, three, etc.) as pronouns. Just as signers introduce a new locus into a conversation, a speaker could introduce a new number into the conversation. Just as signers continue to point to the same location in space to refer back unambiguously to a previously introduced entity, so a speaker could repeat the same number to refer back unambiguously to that entity. The association between number and referent would be no more arbitrary than the association between place and referent, the perceiver can distinguish at least as many numbers as visual places, and the speaker of this language would gain the virtual elimination of referential ambiguity that the signer enjoys. So why does no spoken language do this? Perhaps it has something to do with our associative memories: it may be easier to remember referent/location pairs than to remember referent/word pairs for discourses with three to five participants. Or perhaps the ambiguity that spoken pronouns introduce into a discourse is not a serious enough problem to justify the extra machinery of learning an unbounded list of words the references of which are arbitrary. Either way, it is interesting that sign languages, which have this sort of referential system available essentially for free, always make use of it. Chomsky’s Minimalist Program maintains the idea that noun phrases and pronouns have person, number, and gender features, but eliminates the classical generative notion that they have unique referential indices [Chomsky (1995)]. Sign languages suggest that this might be exactly backwards: when a language has a low-cost stand-in for unique referential indices available (the locus in space), it always opts to use it rather than gender and number distinctions. This suggests to me that the referential index is more basic to the notion of what a (pro)noun is than the so-called phi-features are [convergent with my view in Baker (2003)].

As a final comment, I note that the chapters in this section are fairly typical of the broader literature in that they debate whether different languages have diverse category systems, without having a clear theory of the distinctions among the lexical categories. I think this is a problem that is holding this literature back significantly. In my own work, I realized that I would not be able to tell if the Mohawk language had a distinction between adjectives and verbs until I had a much clearer grasp on what the essential
difference is between adjectives and verbs in languages that clearly have both, such as English. Such a theory cannot be just a list of superficial characteristics, such as that adjectives inflect for case and gender whereas verbs inflect for person, or predicate adjectives need a copula whereas verbs do not. These superficial characteristics vary too much from language to language, and even in languages where they are relevant they cannot be the essence of the matter, providing a unified theory of what it is to be an adjective as opposed to a verb. But if one could get a deeper theory, then it could be applied to languages like Mohawk, and help us get beyond our biases and purely terminological disagreements. For example, Travis’s argument that Agent Raising reveals a difference between adjectives and verbs in Malagasy would be stronger if one could show how that difference follows from a general theory of the adjective–verb distinction – the very same distinction that, when instantiated within a different overall grammar, results in the morphological and syntactic differences that we see in Latin and English. I am not sure how to do this myself; the Agent Raising phenomenon seems to be an unusual and perplexing one. But that should be the ultimate goal. Relatively few generative linguists have attempted to construct a general theory of the category distinctions, but this needs to change. We clearly need to have theories of the syntactic categories that aspire to this kind of explanatory depth in order to finally resolve whether languages differ significantly in their syntactic category systems or not.

References

PART 4

ACQUISITION OF CATEGORIES
This page intentionally left blank
Abstract

This contribution reviews recent research on the acquisition of grammatical categories, focusing on three aspects: parts of speech, inflection, and subcategories of words. It appears that children are sensitive to the distribution of the various classes of words surprisingly early, and that inflectional categories are mastered by age 3 despite the complexity of their acquisition. Subcategories of words take longer to acquire. To account for the facts, it is generally assumed that children are equipped with a distributional learning mechanism, but given the complexity of the factors that must be taken into account, some constraints must be placed on this mechanism. Various types of possible constraints are discussed, from innate knowledge to processing constraints.
1. Grammatical categories

The term “grammatical category” covers a variety of types of categories. We focus here on parts of speech, word classes, and inflectional categories. Part of speech refers to the familiar categories of words such as noun, verb, adjective. In many cases, the words in these general categories split into subcategories with partially distinct grammatical properties. We will call these subcategories word classes. In addition, languages vary with regard to the grammatical or inflectional markers that appear on words (see Table 1 for examples).

Why is it important to know the grammatical category of a word? What does it mean to say that a child “knows” that some element is a “noun”? The answer to the first question is that knowing a word’s category is a precondition for knowing how to use the word in the language. The grammatical category of a word determines (1) the position it is allowed to occupy in the clause (e.g., German verbs appear in final position in subordinate clauses and in second position in main clauses); (2) the range of syntactic functions it can occupy (e.g., a noun may be the subject of a clause, but a preposition or an adjective cannot); (3) the types of words with which it co-occurs (e.g., determiners co-occur with nouns, but not with verbs); (4) the types of morphemes it requires or accepts (e.g., verbs inflect for tense, while nouns inflect for number). How do we know that the missing word in example (1) is a noun? It is not because we can look it up in the dictionary, but because it occupies the typical position of a noun, is modified by an adjective, is preceded by a determiner, and so on. Similarly, it is on the basis of grammatical properties that we know that the word butter in (2) is a verb and not a noun.

(1) The blue ___ is on the table.

(2) The children butter their toast.

<table>
<thead>
<tr>
<th>Parts of speech</th>
<th>Word classes</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>masculine</td>
<td>*le soleil ‘the sun’</td>
</tr>
<tr>
<td>V</td>
<td>feminine</td>
<td>*la lune ‘the moon’</td>
</tr>
<tr>
<td>V</td>
<td>auxiliary</td>
<td>have, will</td>
</tr>
<tr>
<td>V</td>
<td>lexical</td>
<td>stative: *John is resembling Peter</td>
</tr>
<tr>
<td>A</td>
<td>active</td>
<td>active: John is jumping around</td>
</tr>
<tr>
<td>A</td>
<td>attributive</td>
<td>the ball is red</td>
</tr>
<tr>
<td>A</td>
<td>Non-attributive</td>
<td>*the president is former</td>
</tr>
<tr>
<td>Adv</td>
<td>manner</td>
<td>he runs quickly</td>
</tr>
<tr>
<td>Adv</td>
<td>degree</td>
<td>*he runs very</td>
</tr>
<tr>
<td>D</td>
<td>definite</td>
<td>the ball</td>
</tr>
<tr>
<td>D</td>
<td>indefinite</td>
<td>a ball</td>
</tr>
<tr>
<td>P</td>
<td>contentful</td>
<td>around</td>
</tr>
<tr>
<td>P</td>
<td>grammatical</td>
<td>of</td>
</tr>
</tbody>
</table>

Table 1
Parts of speech and word classes
Thus, to say that a child knows that some word is a “noun” is to say that he or she uses this word in the typical sentence positions occupied by nouns, with the proper inflections, and so on. To produce grammatical sentences in the language they are acquiring, children must use the words according to the properties determined by their category. The specific label attached to a part of speech is not what interests us here. What matters is the fact that the different parts of speech have different grammatical properties. This is clearly expressed by Pinker (1984, p. 43): “The only significance of the name of a symbol (like Noun) in a cognitive account is that the process that manipulates such symbols treats all the symbols with a given name alike but differently from symbols with different names.”

The child learning a language must categorize every new word according to its part of speech, and in some cases according to some particular word class within that part of speech. The child must also identify the inflections on words and categorize them as denoting gender, number, grammatical role, or any of the possible inflectional categories. Moreover, he or she must assign each word to the proper inflectional class, as it is frequently the case that the phonological form of inflectional markers is not uniform across the language. For example, nominative plural ending in German is not the same for all masculine nouns, as illustrated in Table 2.

We will start with a discussion of parts of speech, then move on to inflectional categories and finally to word classes.

### 2. Two-word utterances and their analysis

Let us consider some typical two-word combinations produced by English-speaking children around their second birthday, and some characterizations of them, shown in Table 3. The grammatical description in the second column assumes that young

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Inflectional classes: German masculine nouns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Der Tag (the day)</td>
</tr>
<tr>
<td>Nominative Plural</td>
<td>Die Tage</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Some two-word utterances and their possible analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child utterance</td>
<td>Part of speech description</td>
</tr>
<tr>
<td>All fix</td>
<td>Adv + V</td>
</tr>
<tr>
<td>See baby</td>
<td>V + N</td>
</tr>
<tr>
<td>More cookie</td>
<td>Adv + N</td>
</tr>
<tr>
<td>Boot off</td>
<td>N + Part.</td>
</tr>
<tr>
<td>Mommy shoe</td>
<td>N + N</td>
</tr>
</tbody>
</table>
children categorize words into the part of speech categories of the adult language. Under the semantic approach typical of the 1970s (third column), children’s first word combinations are the result of rules mapping over semantic categories [Schlesinger (1971), Bowerman (1973)]. In Braine’s Pivot Grammar [Braine (1963)], fourth column, the grammar of young children contains only two categories of words: Pivot words and Open words. The latter two approaches must explain how and when children abandon the initial grammar and move on to a grammar based on part of speech categories, recategorizing every word in terms of the adult grammar in the process. As we will see, recent evidence shows that children attend to the formal properties of words from the beginning.

Table 3 also illustrates the fact that the utterances produced by English-speaking children around 24 months of age are devoid of functional elements like determiners, auxiliaries, and morphological endings. Some authors have suggested that children may not attend to these elements, presumably because they are unstressed [Echols and Newport (1992)]. Their internalized grammar would not yet contain functional categories [Radford (1995, 1996)]. Here again, we will see that children take functional elements into account long before they display knowledge of these categories in their production.

3. A semantic approach to grammatical categorization: Semantic bootstrapping

How does a child learn to categorize words as being verbs, nouns, adjectives, or adverbs? A view that came to be known by the name semantic bootstrapping is based on the idea that semantics guides children into categorizing the words of their language: a word denoting an object or a person is a noun, a word denoting an event is a verb, and so on. Children would use semantics to “pull themselves up by their bootstraps” and master syntax. There are two main approaches to semantic bootstrapping.

A cognitive approach assumes that infants organize the world into cognitive categories such as objects, events, and so on. When they start learning language, they initially classify the words they learn in terms of these cognitive categories [Brown (1973), Macnamara (1982)]. Grammatical categories like “noun” or “verb” emerge when children start learning the properties of words.

The innatist hypothesis [Grimshaw (1981), Pinker (1984)] assumes that the child is innately equipped (1) with two types of universal knowledge – a set of part-of-speech categories and a set of semantic categories – and (2) with a universal default mapping between the two, as shown in Table 4. The child infers the categorization of words on the basis of their semantic properties.

Semantic criteria do not provide an adequate basis for determining the part of speech of a word. Many nouns do not refer to physical objects; some denote actions (a hug), events (an explosion), or states (a depression). Many verbs do not refer to actions (want); some denote a property (resemble). Some adjectives are actional (quick, fast) or controllable (be polite!) or denote concepts expressed in other languages with verbs
(hungry, wet, sleepy), and so on [Maratsos and Chalkley (1980), Maratsos (1999, 206)]. If children start out with semantic categorization criteria, at some point they must abandon them; otherwise they will not arrive at the proper grammar. In both approaches to semantic bootstrapping, only the first set of words in each category is semantically classified. Once some elements have been classified, the child starts analyzing the grammatical properties of these words, and uses the distribution of items that do not conform to the basic semantic type to posit their category on the basis of known structure. Gradually, grammatical properties come to predominate as classifying devices [Macnamara (1982)].

Semantic bootstrapping has been criticized both on theoretical and on empirical grounds. First, Benedict (1979) classified the first 50 words understood and produced by eight children according to semantic criteria based on the children’s use of these words. Her class of “action words” contained interjections (peekaboo), verbs (eat, give), adverbs (no), and particles (out, up, down), while her class of “modifiers” contained adjectives (big, hot), adverbs (there), pronouns (mine), and unanalyzed phrases (allgone). If children classified all these words as verbs in the first case, and as adjectives in the second, they would be in deep trouble. Yet, there is no evidence that children misclassify words to a great extent [Maratsos and Chalkley (1980), Braine (1987)]. This suggests that children attend from the beginning to distributional information.

Second, semantic bootstrapping would be supported if it could be shown that the first nouns that children learn designate objects or people and their first verbs, actions. But Bassano (2000), in a study of the first words produced by a French-speaking child, observes that concrete action verbs are not the earliest verbs to be produced with any frequency; situational, attention-getting verbs, modals and être ‘be’ are the first to occur. She also reports that neither nonconcrete object names nor nonconcrete action verbs are avoided in the child’s early speech.

Third, the innatist approach to semantic bootstrapping crucially assumes that there is a universal set of grammatical categories. Many typologists would reject this assumption [Culicover (1999, p. 39)]. For example, in Japanese, there are two categories of adjectives (two grammatically distinct sets of words denoting properties), while in other languages there are no adjectives, and adjectival meanings are expressed by nouns (words denoting properties are non-distinct grammatically from words

<table>
<thead>
<tr>
<th>Grammatical element</th>
<th>Semantic inductive basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>Name of person or thing</td>
</tr>
<tr>
<td>Verb</td>
<td>Action or change of state</td>
</tr>
<tr>
<td>Adjective</td>
<td>Attribute</td>
</tr>
<tr>
<td>Preposition</td>
<td>Spatial relation, path, or direction</td>
</tr>
</tbody>
</table>

Table 4
Syntax-semantics correspondences according to Pinker (1984)
denoting entities), as in example (3), or by verbs, as in (4) [Schachter (1985, p. 17, ex. 32, 33, p. 18; ex. 40, 41)].

(3) a. **Rikashka: hatun-kuna-ta** (Quechua)
   I-saw big-pl-acc
   ‘I saw the big ones.’

   b. **Rikashka: alkalde-kuna-ta** (Quechua)
   I-saw mayor-pl-acc
   ‘I saw the mayors.’

(4) a. **Piaoliang de nühaizi** (Mandarin Chinese)
   beautiful rel girl
   ‘a girl who is beautiful, a beautiful girl’

   b. **Liaojie de nühaizi** (Mandarin Chinese)
   understand rel girl
   ‘a girl who understands, an understanding girl’

Even such a basic distinction as that between noun and verb does not appear to be universal. In **multicategorical** languages like Tagalog, entity-denoting roots and event-denoting roots seem to have identical grammatical properties; both may function as predicates or as arguments [Schachter (1985, p. 11)].

(5) a. **Nagtatrabaho ang lalaki** (Tagalog)
   is-working TOP man
   ‘The man is working.’

   b. **Lalaki ang nagtatrabaho**
   man TOP is-working
   ‘The one who is working is a man.’

**Multicategoriality** poses a major problem to a semantic bootstrapping approach. This is seen more clearly if we consider the multicategoriality of some frequent English words: *kiss* is a noun in *Give me a kiss* but a verb in *Kiss me* (as are *hug, drink, bite, call*, etc.) [Nelson (1995)]. If children interpret *kiss* as an action word, and assume that it is a verb, and if they hear *kiss* sometimes in a verb frame and sometimes in a noun frame, how are they to infer the grammatical properties of verbs and nouns? In languages like Tagalog, where multicategoriality is generalized to the whole lexicon, the syntax of the language cannot be learned via semantic bootstrapping.

We conclude that, while semantic categories certainly participate in word classification, in that if a child notices that a word denoting an action has property $x$, he or she might expect other words denoting an action to have the same property, these categories do not form the initial basis for grammatical categorization. To arrive at the adult grammar, children must register the distributional properties of lexical items.
4. Distributional learning

A number of computer simulations have shown that distributional learning can go a long way in categorizing the words of the language on the basis of positional and co-occurrence information [e.g., Brent (1994, 1996), Cartwright and Brent (1997)]. Also, studies of caretakers’ language show that distributional information is a reliable cue to grammatical categories in the language children hear [e.g., Mintz, Newport and Bever (2002)]. Moreover, psychological experiments with artificial languages demonstrate that not only adult but also infant learners are able to perform distributional analyses [e.g., Braine et al. (1990), Brooks et al. (1993), Mintz (2002), Saffran, Aslin and Newport (1996), Valian and Coulson (1988)]. Thus, infants have the capacities and the proper input to learn grammatical categories on the basis of distributional information. In this section, we discuss evidence showing that children register formal cues in the language and use them to categorize words. We distinguish three types of formal cues: word order, morphological cues, and co-occurrence restrictions.

4.1. Word order

Studies of comprehension indicate an early sensitivity to word order. Using a preferential looking paradigm, Hirsch-Pasek and Golinkoff (1996) show that infants as young as 16 months old can distinguish “Big Bird is tickling Cookie Monster” from “Cookie Monster is tickling Big Bird”: upon hearing the first sentence, they tend to look longer at a video showing Big Bird doing the action (as opposed to a video where Cookie Monster is doing the action).

In language production, the earliest moment when we can start observing word order effects is when children start putting two words together. As long as children produce only one word at a time, we cannot talk of word order. Do the first two-word utterances produced by young children already show a word order effect? Yes. Children produce utterances like: more juice, more car, more read, more hot, with more in first position, and boot off, hat off, shoe off, with off in second position. These two-word utterances seem to be the product of highly restricted limited-scope formulae containing a relational lexical item (more, off) and a positionally specified empty slot for its argument: more+X, X-off [Braine (1976)]. The consistent word order shows that children register recurring combinations in the input, and are attentive to positional cues.

Children seem to treat nouns and verbs differently almost from the beginning. In English, many verbs alternate between a transitive use and an intransitive use, where the subject of the intransitive corresponds to the object of the transitive:

(6) a. John broke the vase.
   b. The vase broke.

The alternation can be described as a rule allowing a verb to switch between two argument frames. Tomasello et al. (1997) taught novel verbs and novel nouns (e.g., wug) to
children aged between 1;6 and 1;11. The words were taught in one frame, and the authors looked at whether the children could generalize to other frames. They found that children were conservative with novel verbs, which they tended to use only in the type of frame in which they were taught, reproducing the modeled word order. But when taught a novel noun, they could instantly use it in the two-word combinations that they had mastered, either in first or in second position (wug gone, more wug). The authors suggest that children of that age possess some kind of paradigmatic category “noun” corresponding to the type of item that can fill the variable slot in the limited-scope formulae that they know, but not the category “verb” [Tomasello and Brooks (1999, p. 168)]. Fisher (2002) argues that the difference in the treatment of nouns and verbs is due to the different syntactic role that verbs play in clauses: verbs are relational elements. Naigles (2002, in press) also points out that children may be reluctant to generalize verbs across frames because a change of frame implies a change of perspective on the event, whereas this is not the case with nouns. In comprehension, children seem to be better able to switch from one frame to another [Naigles and Hoff-Ginsberg (1995)].

This research shows that children distinguish relational terms from non-relational terms. When learning names for entities, they encode only the type of referent, but when learning relational terms like verbs or words like more and off, they encode both the approximate meaning of the term and its argument frame, more or less as in (7).

(7) more: meaning: recurrence of X
    argument frame: more + X

In (7), the position of the argument is specified, but not its grammatical category (“noun”). Children produce not only utterances like more cookie, with a noun as complement of more, but also more read ‘read some more’ or more hot ‘another hot thing’ [Braine (1976)]. This shows that the fully specified grammar in column 2 of Table 3 attributes too much knowledge to the child1. In the early grammar, the argument frame is probably underspecified with regard to the grammatical category of the complement.

The absence of generalization with verbs ties in with other research showing that children learn constructions lexical item by lexical item and do not generalize from one lexical item to another [Braine (1976), Peters (1983), Tomasello (1992)]. Indeed Akhtar and Tomasello (1997) have shown that when children younger than 3;0 hear a novel transitive verb for which the argument structure has not been modeled (“This is called dacking”), they are unable to use or comprehend word order to mark agents and patients. One is reminded of Karmiloff-Smith’s (1986) model according to which the initial phase of learning is a list of independent procedures; these procedures are integrated into more general ones at a later stage involving representational redescription.

The conceptual simplicity of nouns as labels for objects compared to the more complex role of verbs as relational elements has been proposed [Gentner (1982)] to account for the predominance and earlier acquisition of nouns over verbs in English, Italian,

---

1 For the same reason, the open slot is not specified in semantic terms, such as more + substance, contrary to the semantic approach illustrated in Table 3.
Spanish, and French [Bassano (2000), Bates et al. (1994), Caselli, Casadio and Bates (1999), Jackson-Maldonado et al. (1993)]. This is the so-called “noun bias” in acquisition. As the noun bias is not observed in the acquisition of Asian languages [Choi (1997), Choi and Gopnik (1995), Gopnik, Choi and Baumberger (1996)], it seems that language-specific factors are at play; for example, Asian languages allow massive argument deletion, resulting in many utterances containing only a verb [de Boysson-Bardies (1996)].

4.2. Inflection and inflectional class

Much of the information necessary to determine the grammatical category of a word is coded by function words like determiners or auxiliaries, and by inflectional morphemes. Thus, upon hearing the nonsense clause *The wug zaks*, we know that *wug* is a noun (because it is preceded by *the*) and *zaks* a verb (because it agrees with *wug*). Valian and Coulson (1988) have shown that frequent markers are anchor points for distributional analysis: adults have more success in learning an artificial language when the language contains frequent markers. When do children start registering function words and inflection? Do they use this information to categorize words? We might hypothesize, for example, that a child learning English would observe *-ed* on certain words and use this cue to determine membership in the category Verb. While this appears trivial, it is not. Affixes (prefixes or suffixes) can be inflectional, but they can also be category-changing: adding *-ly* to adjectives turns them into adverbs. As stressed by Culicover (1999, p. 38): “The formal type of knowledge can only be brought into play after the basic categories are established, at least in some preliminary way. Once a learner knows that a particular formal inflection is a property of a certain category, this information can be used to categorize new elements.” In order to use an affix as a cue to the category of a word, the child must first identify the affix and second determine which type of word it appears on. But an inflectional morpheme is also a grammatical element that must be categorized as marking some grammatical function. In this section, we focus on inflectional morphology as a categorization problem; in Section 4.3, we turn to the question of whether children use function words and inflectional morphology to determine the category of words.

Categorizing an inflectional morpheme as marking some grammatical function is not a simple matter, since the marker-to-function mapping is often many-to-one or one-to-many. A case of one function/many forms is observed in declension classes. In inflected languages, it is frequently the case that a part-of-speech category is subdivided into subclasses, each one accepting a specific set of forms (see Table 2 for German masculine nouns in the nominative plural). In Latin, there were five declension classes. In Bantu languages, there may be up to 20 different word classes indicated by obligatory prefixes [Suzman (1996)]. Such classes are in general arbitrary without any semantic core to them [Braine (1987)]. In addition, in many cases, the pronunciation of an affix varies according to its phonological environment, and children must determine that the various forms they hear are grammatical variants of the same underlying morpheme [Peters (1997)]. In some cases, the phonological changes brought about by a morpheme are so complex that it is difficult to identify the derived word as being fundamentally
“the same word” as the initial one. [For example, in Chickasaw, hilhali ‘I’m dancing’ becomes akhi’lho ‘I’m not dancing’ in the negative: Anderson (1985, p. 165)].

Homonymy leads to a situation of **one marker/many functions**. To give some examples, in Polish, the suffix \(-a\), which marks nominative case in class 1 nouns, marks genitive case in class 2 nouns [Maratsos (1999)]. The clue provided by the affix can be relied on only after the learner has some knowledge of the declension classes. In English, \(-s\) is not a reliable cue to the category Verb, as it marks verbs for third person singular, nouns for plural, and is also a possessive marker.

An example provided by Maratsos (1998, 1999) is worth citing in full. In Turkish, functions are marked by inflectional case markers. The suffix \(-u\) on a noun means that it is the patient of the action (the direct object). How can the child “notice” that \(-u\) marks patienthood? Since word order is quite free, the child cannot rely on that. Suppose the child hears the Turkish equivalent of “Sam \(-u\) scratched Ann” and can tell from the context that Sam is the patient. Is that sufficient to learn that \(-u\) marks patienthood? No, says Maratsos, because \(-u\) could mark many other things: gender (masculine), the word-class of Sam (long, thin object), humanness, animacy, third person, social status, affectionate regard on the part of the speaker, or any other category languages mark on nouns. To determine the value of the suffix, the child must be able to register the properties actually encoded by the language. If we add to that the fact that, in Turkish, not all patients are marked with \(-u\), only definite patients, and that there are actually four variants of the \(-u\) suffix ([u], [ü], [i] and [ï]) chosen according to the phonological characteristics of the stem, so that no single form consistently marks patients, we can see that learning the value of the suffix \(-u\) is not an easy matter, because this form appears on only a small subset of nouns denoting patients, perhaps 10–15%. A serial induction process in which the child guesses at the property encoded by the form and tests the validity of these guesses against the input, one at a time, until one is found that works [Pinker (1984)], would have to have a very low threshold of success to be applicable. On the other hand, if the child registers all possibilities at once and then checks off the ones that do not match upon subsequent exposure to the suffix, he or she would have to consider not only the possible values of the form, but also the phonological characteristics of the stem and the definiteness of the patient. Maratsos concludes that the child cannot just “notice” the value of inflectional morphology upon a single exposure; he or she must “grind through” a large number of possibilities over a wide variety of utterances.

When we also consider that a single affix may convey more than one piece of information (the ending \(-ai\) on the French verb *marchai* ‘walked’ marks both tense – Passé Simple–and person-number of the subject – first person singular); and that in highly inflected languages like Greenlandic Eskimo (8), roots are constructed with more than one affix, each one requiring identification and categorization, we have to conclude that learning inflection is not an easy matter.

(8) uppi-ti- le- qa- akkit
    fall- cause-begin-intensifier-1st/2nd singular indicative
    ‘I’m going to make you fall!’
How do children cope with such difficulties? Surprisingly, perhaps, crosslinguistic studies show that children make very few errors. When they use inflectional morphology, they use it correctly; errors are more often of omission than of commission [Phillips (1995), Maratsos (1998)]. Moreover, children master the essentials of the inflectional system of their language by their third birthday, provided that the system is regular and phonologically transparent. The case systems of Turkish and Polish are mastered with no observed errors by age two [Aksu-Koc and Slobin (1985), Smoczynska (1985)]. Children learning highly inflected languages produce inflectional markers productively much earlier than English-speaking children; example (8) was produced by a 2-year-old [Forteacute;scue and Lennert Olsen (1992)] [see also Clancy (1985), Choi (1997)]. The complex word-class system of Bantu nouns seems to be mastered with almost 100% accuracy by 2;6 [Suzman (1996)]. When inflectional paradigms are phonologically complex, riddled with homophony, semantically opaque, or inconsistent, children master the essentials of the system by their third birthday, but make errors with exceptional forms, which they tend to regularize.

The robustness and efficiency of the acquisition of inflectional morphology suggest that some innate factors are at work, faculty-specific, species-specific, or both [Maratsos (1998)]. As the conceptual domains which are grammaticized in languages seem to form a closed list, it is tempting to think that learning would be helped if children limited their search to that list. The child might also be helped by an innate knowledge of the universal correspondence between certain grammatical categories and certain types of markers (see Table 5). For instance, once the child has identified a suffix indicating tense, he would automatically know that the word it is suffixed to is a verb.

But this strategy would not work in every case. In Nootka, the inflectional markers for tense and mood attach to the right of the first word in the sentence; in (9) this word happens to be the accusative case marker of the noun phrase ‘the deer’ ?ooqw bowatc ?aq [Anderson (1985, p. 156, ex. 1a-b)]. In Kwakw’ala (example 10), grammatical particles attach phonologically to the preceding word rather than to the word they determine [Anderson (1985, p. 166, ex. 2)].

(9) ?ooqw-obt-qa bowatc ?aq tl’itcitl John
   acc -past-declar deer det shoot John
   ‘The deer, John shot (it).’

<table>
<thead>
<tr>
<th>Category</th>
<th>Typical markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>person, number, tense, aspect, mood, potential, desiderative, causation, voice, subject and object agreement, speech act type, social status of interlocutors, speaker’s evidence for making claim, etc.</td>
</tr>
<tr>
<td>Noun</td>
<td>number, gender or word class, case, definiteness, animacy, form, orientation, location with respect to speaker, etc.</td>
</tr>
</tbody>
</table>

Table 5
Typical markers accompanying lexical categories [based on Slobin (1997, 2001)]
While innate knowledge constraining the search space would help, we must concur with Maratsos that children must “grind through” the data to discover the correct generalizations.

As for the idea dear to cognitivists that the acquisition of grammatical markers is driven by the child’s need to express some cognitive function, Bowerman (1985) argues that this cannot be the case. The acquisition of determiners in French or English is not driven by the child’s desire to communicate whether a noun is already known to the speaker – definite (the dog) – or not – indefinite (a dog) – but by the necessity, grammaticized in this language, of encoding it. We are reluctant to recognize this because familiarity makes us take for granted the distinctions encoded in our own language, but, says Bowerman,

consider the obligatory four-way classification of nouns in sentences of Toba, a language of Argentina, according to whether the objects to which they refer are in view, out of view, coming into view, or going out of view, and furthermore, if they are in view, according to whether they are spatially non-extended (e.g. a fruit), extended vertically (e.g., a fruit still hanging, or a tree), or extended horizontally (e.g., a table) … Can such meanings really struggle for expression in the developing minds of all children, including those in our own living rooms?


In fact, one recurrent observation in the language acquisition literature is that formal distinctions orient the child toward discovering the semantic relations they encode, rather than cognitive distinctions orienting the child toward finding the formal way to express them [Bowerman and Choi (2001), Slobin (2001)]. Phonology, in particular, plays a crucial role in the acquisition of morphology. Children learning Bantu languages rely, not on semantics, but on phonological information to learn the complex noun class system [Demuth (1992), Suzman (1996)]. Demuth suggests that this holds crosslinguistically: “Access to the semantics of the system becomes available only at later stages of development, whereas early overgeneralizations are normally of a phonological nature” [Demuth (1992, p. 630)]. The same point is made by Naigles (2002), who argues that children find form easy, but pairing of form and meaning hard, in essence because it requires more computational resources [see also Peters (1997)].

If children must discover the function of morphemes by observing the way they are used in the input, we might expect them to entertain erroneous hypotheses for a while. This is discussed by Eve Clark (2001) under the term “emergent categories.” Clark suggests that children may briefly use a morpheme of their language to express distinctions made in other languages, but not in their own. She gives the example of a child who encoded the difference between inherent and temporary properties by using respectively adjectives in -y (It’s crumby = full of crumbs, said of a biscuit; D 2; 6,9) and adjectives in -ed (My foot is all crumbed = covered with crumbs; D 2; 6;30) (see also Clark, this volume). Again, this raises the question: are there built-in constraints as to the kinds of hypotheses children might entertain?
4.3. Function words

We have seen that children attend to inflectional morphology from the beginning. Another cue to the category of a word is provided by the types of function words with which it co-occurs. Do children register the function words in sentences, and do they use them to determine the grammatical category of content words? Children learning English might, for example, notice that words denoting concrete objects are preceded by a determiner like a or the. Even though not every concrete object word is immediately preceded by a determiner (a big truck) and not every word preceded by a determiner denotes a concrete object (a hug), they could use the statistical regularity to infer that a novel word preceded by a determiner (a wug) might denote a concrete object.

In a seminal study, Brown [(1957), replicated in Dockrell and McShane (1990)] taught a novel word to 16 children 3 to 5 years old, while presenting them with pictures of some novel activity performed on some confetti-like material. Three different formulations were used. In the verb frame, the question was: “Do you know what it means to sib? In this picture, you can see sibbing. Now show me another picture of sibbing.” In the count noun frame, the question was: “Do you know what a sib is?” And in the mass noun frame, it was: “Have you ever seen any sib?” When the new word was introduced as a verb, the children tended to choose the picture displaying the same action (10/16); with the count noun frame, they chose the picture displaying the same object (11/16); and with the mass noun frame, they chose the similar substance (12/16). Thus, 3- to 5-year-olds can infer the part of speech of a word on the basis of its structural environment and use this information to determine the meaning of words.

Function words – functors – like determiners and auxiliaries are not only cues to the grammatical category of other words; they are themselves grammatical elements which must be categorized. They are characterized by a cluster of phonological properties such as lack of stress, monosyllabicity, short syllable duration, null coda, and vowel harmony, which distinguish them from full lexical items [Shi, Morgan and Allopenna (1998)], but which also make them less salient. When do children start taking these elements into account? Within the generative framework, functional elements head syntactic phrases, building a functional structure above lexical items. A central theoretical problem is exactly how much functional structure children master. The hypotheses range from a full competence approach, whereby all functional categories are assumed to be present from the start [Poeppel and Wexler (1993), Borer and Rohrbacher (1997)], to structure building, whereby the functional structure is built up as children acquire the functional elements [Radford (1995, 1996)]. [For a discussion, see Guasti (2002)]. In the latter case, the appearance of filler phonemes (the proto-determiners /i/, /a/, /n/ in (11)) where a functional morpheme is expected would denote the beginning of an awareness of the presence of the corresponding functional category [Peters (1995), p. 472]):

(11) /i ká:/ ‘F car’ (F = filler)
/a gÚdkí/ ‘F cookie’
/n báp/ ‘F bump’
The structure building hypothesis seems to account for the gradual accretion of functional markers in language production, but it is confronted with the problem of explaining the difference between comprehension and production. A child who does not produce some functional morpheme should not be able to use it in comprehension as a cue to the grammatical category of words. But a growing number of studies shows that functional elements are taken into account long before children start producing them. For example, 24-month-old single-word talkers react negatively to sentences containing incorrectly used grammatical morphemes or where the grammatical morpheme is omitted [Golinkoff, Hirsch-Pasck and Schweisguth (2001)]. At a stage when they do not produce grammatical morphemes, children carry out fewer commands when the grammatical morphemes are omitted (Throw ball vs. Throw the ball) [Shipley, Smith and Gleitman (1969)]. When asked to imitate sentences containing real or pseudo-functors, children with a mean length of utterance (MLU) lower than two words leave out the real morphemes more often than the pseudo morphemes [Gerken, Landau and Remez (1990)]. Children confronted with grammatical stimuli of the type Find the bird for me and ungrammatical stimuli with a pseudo-functor Find gub bird for me or an erroneous functor Find was bird for me respond correctly to the grammatical stimuli more often, showing that they are aware of where the real functors occur and what they sound like [Gerken and McIntosh (1993)]. Even children with MLUs of less than 1.5 words who spontaneously produced no determiners at all responded significantly better to commands with grammatical morphemes. Already by 16 months, children have an idea of where function words are likely to occur, given that they show a listening preference for natural passages, over passages where function words have been misplaced [Jusczyk (2001)]. Finally, infants as young as 10.5 months distinguish normal English passages from passages where nonsense syllables replace function words [Shady (1996)].

Such studies show that very young children notice grammatical elements and are sensitive to their distribution. But are they able to use morphology and function words to determine a word’s part of speech? Waxman and Booth (2001) showed that, when presented with a novel word in a noun frame (These are blickets; this one is a blicket) or in an adjective frame (These are blickish; this one is blickish), children aged 14 months with a mean production vocabulary of 15 words map nouns to categories of objects. They attend to the cue provided by the determiner or the plural morpheme up to a year before they display these elements in their production. The adjective frame did not consistently elicit a property interpretation at 14 months, but preliminary evidence suggests that it did by 24 months².

To sum up, very young children are aware of the morphology appearing on lexical items and of the position and form of function words, and they are able to use this knowledge to determine the grammatical category of words and their interpretation.

² The main cue for the property interpretation in the adjective frame used is the infrequent ending -ish on the word. The syntactic position of blickish after the copula in This one is blickish is the same as that of proper names (This one is Fred), and of mass nouns (This is sugar), That could explain the results with the adjective frame.
Just how much knowledge of function words children have at this early stage of language learning is a question that we are far from being able to answer at this point. Why is it that children seem to understand so much, yet produce so little? It could be that children do not know enough about the function or meaning of functional morphemes to be able to integrate them in their productive grammar. For example, children might have categorized the, a, and an as belonging to the category D of determiners, and might use this knowledge to determine that the following word is a noun, while they might not have attached features like [+definite], [+singular] or [before Vowel] to them. Since such features govern which determiner to choose in a given context, children do not have enough knowledge of these words to use them. Observe that the proto-determiner /n/ in (11), presumably linked to an, is used in front of a word beginning with a consonant. The child has an idea of what determiners sound like but he doesn’t quite know which form to choose in a given context.

Gerken [(2001) and references therein] argues that the omission of grammatical morphemes in the language of young children is to be attributed to a metrical constraint on production. Children’s utterances would tend to be limited to sequences of weak syllables alternating with strong syllables, and grammatical morphemes would be omitted when they do not fit into this pattern. Children are likely to say *He hugs the dog* but *He kisses dog*, omitting the determiner because the weak syllable is taken up by the plural -(s)es. This phonological approach is not incompatible with the incomplete knowledge approach sketched above, and might well complement it.

4.4. Word classes

Children must not only determine whether a word is a noun or not, but also, if it is a noun, which type (subcategory) of noun it is. English distinguishes between count nouns (*car*), mass nouns (*sand*), and proper nouns (*Peter*). Count nouns occur with determiners like *a*, *many*, *several*, mass nouns follow determiners like *some* and *much*, and proper nouns are not preceded by determiners.

Here again, a number of studies have shown that children are aware of the distinction surprisingly early. In a classic study, Katz, Baker and Macnamara (1974) presented 17-month-old children with a novel word for a doll. For some children, the noun was preceded by a determiner (*This is a wug*), for others, it was not (*This is wug*). The children tended to interpret the word preceded by a determiner as naming the kind of doll, and the bare noun as giving the name of the specific doll. This distinctive behavior was observed mainly with girls, and it occurred only when the object labeled was a doll (i.e., person-like), and not when it was a block. This study taps into the beginnings of children’s capacity to take into account the presence or absence of the determiner, and to use it as an indication that a word is a common noun or a proper noun [see also Gelman and Taylor (1984)]. Soja (1992) got similar results with the distinction between count and mass nouns with 2-year-olds. Such experiments show not only that children categorize nouns on the basis of the cue provided by the determiner, but also that they use what they know of the denotation of a mass noun or a count noun to infer the meaning
of words. Thus distributional cues are actively used by children in interpreting sentences [a strategy called “syntactic bootstrapping” by Gleitman (1990), in the context of the acquisition of verb meanings]. But we should guard against attributing too much competence to the children: when form class cues conflict with word learning constraints like the whole-object assumption, children often ignore the formal cues [Woodward and Markman (1998)].

While the subcategories of nouns are learned early, other word classes are late acquisitions. In some cases, learning a subcategory of words requires establishing a dependency between words. For example, a consistent crosslinguistic finding is that regular pronouns (him) are learned later than anaphors (himself). Preschool children know that in John hurts himself, himself must refer to the subject John and not to someone else; but they think that in John hurts him, him may also refer to John [see Thornton and Wexler (1999), Guasti (2002), for discussion and references]. Here the subclass of a pronoun determines its interpretive possibilities with respect to a potential antecedent within the clause. In order to learn the distinction, children must consider dependencies between non-adjacent words.

Verbs are also subcategorized into various classes, such as transitive and intransitive. One intriguing problem is that of the transitivity alternations illustrated in (6), and repeated below.

(12) a. John broke the vase.

b. The vase broke.

The alternation is limited to certain subclasses of verbs. Verbs of manner of motion alternate (swing, slip), but verbs of directed motion do not (some are intransitive: rise, fall; others are transitive: pull, raise). How do children learn these narrow conflation classes? Pinker (1989, p. 270) argues that traditional category formation would not work because there is no simple definition that would include only the classes of verbs that alternate and exclude those that do not. He proposes to equip the child with innate knowledge of the types of semantic structures available for verbs; upon learning that a verb alternates, the learner would assume that other verbs with the same semantic structure alternate. Braine and Brooks (1995), on the other hand, propose that children adhere to a “unique argument-structure preference” principle: once one argument structure is firmly learned for a verb, it tends to preempt other argument structures for that verb until the language they hear teaches them otherwise. This view, which would explain the rarity of transitivity errors (*I’m falling it), is supported by a study by Brooks and Tomasello (1999), who taught children novel verbs in one construction and looked at whether they were ready to use them in the other frame. Children aged 2;5 tend to use verbs only in the construction in which they are learned. When they produce transitivity alternations, they do not distinguish between alternating and non-alternating verb classes. It is only by six or seven years of age that the narrow classes seem to be acquired.

Culicover (1999) argues that narrow range classes of words are ubiquitous in language. For example, only a small subset of adjectives can appear in the impersonal construction
in (13), and, of this set, a tiny subset can also appear with a similar meaning in the construction in (14) [Culicover 1999, p. 47).

(13) It is likely that Robin will be elected President next year. 
probable, possible, certain, clear, sure

(14) Robin is likely to be elected President next year. 
*probable, *possible, certain, *clear, sure

How do children manage to learn these tiny subclasses? According to Culicover, the learner must be a conservative attentive learner, who will recognize regularities, but will generalize only if the number and distribution of cases exceeds certain bounds [Culicover (1999, p. 29–30)]. Because of the variety of minor formal categories that must be learned, because distributional mechanisms are necessary to learn them, and because the same mechanisms can construct major formal categories, Maratsos (1998, p. 447) concludes that to hypothesize innate knowledge of major categories is not only unnecessary, it is also theoretically unparsimonious. It could, however, be claimed that the only way children can home in on the proper subclasses is if they have innate knowledge of the grammatical factors relevant for distinguishing them.

4.5. Other cues to grammatical category learning

It has been suggested that children might exploit phonetic or phonological cues to learn parts of speech. In English, nouns are more likely than verbs to be stressed on their initial syllable (récordN vs. recórdV). If children notice this regularity, they could exploit it to categorize new words [Kelly (1996), Gerken (2001, 156)]. In the case of English, it is not clear that this strategy would be much help, because by the time the child has learned enough words to notice the regularity, he or she is probably able to categorize nouns and verbs on the basis of distributional information alone. But this strategy cannot be discounted. Peters (1997) discusses various other phonological and prosodic factors influencing the acquisition of grammar, in particular stress, saliency, and rhythm [see also Gleitman and Wanner (1982), Morgan (1986), Jusczyk (1998, 2001)].

5. Models of distributional learning

The evidence reviewed thus far shows that children are sensitive to distributional information. Before the age of two, they attend to morphosyntactic elements and use them as cues to determine the grammatical categories of words. In this section, we review some distributional learning mechanisms that have been proposed to account for the acquisition of grammatical categories.

An early model of a distributional learning mechanism is that of Maratsos and Chalkley [(1980), Maratsos (1982)]. The child registers the semantic, morphological, and distributional properties of lexical items. Once a pattern has been established with one lexical item, other lexical items can assimilate to this pattern, and other patterns can assimilate to
this lexical item. As more and more lexical items are learned, the recurring properties of grammatical classes emerge. In Bates and MacWhinney’s (1987) Competition model, children attempt to map grammatical forms with their semantic function by taking semantic, morphological, and positional cues into account. In Brent’s (1996) autonomous bootstrapping model, children extract a tiny bit of linguistic knowledge in any domain from unanalyzed inputs, and use that bit of knowledge to perform a little more linguistic analysis on future inputs, thereby extracting more knowledge from them [Brent (1996, p. 25), Cartwright and Brent (1997)]. As more and more knowledge is extracted, the number of cues that children can use to parse incoming input grows exponentially.

Important notions for cue-based learning models are cue validity – the availability and reliability of some cue in the input (an objective notion) – cue strength – the weight that the organism attaches to some piece of information (a psychological notion) – and cue cost – the cost of processing a cue [Kail (2000)]. The cues do not need to have 100% validity. Learning depends on discovering new cues and modifying the strength of known ones. The various cues are probably learned independently of each other, each with its own representational content [Rispoli (1999, p. 234)]. They may reinforce each other or, in some cases, compete with each other [Dockrell and McShane (1990)].

Neural (connectionist) networks can be viewed as cue-based distributional learning mechanisms that compute cue strength as a function of cue validity. They basically perform pattern association. Connectionist models can be quite successful at learning various aspects of language [Rumelhart and McClelland (1987), Plunkett (1995)], but Pinker has repeatedly argued that pattern association is not sufficient to account for grammar learning [Pinker (1999), Pinker and Mehler (1989), Pinker and Ullman (2002)]. Pinker and Ullman (2002) and Palmer-Brown, Tepper and Powell (2002) also claim that recent connectionist models are really hybrid models of parsing, where some learning is associative, but structure is built into the model [see also Karmiloff-Smith (1992), Elman et al. (1996), Karmiloff-Smith et al. (1998)]. In this context, one interesting hypothesis would be that a connectionist-type distributional learning mechanism classifies new information and categorizes it with similar elements, and that at some point, an internal process of representational redescription [Karmiloff-Smith (1986)] extracts the general rule from the knowledge base and stores it separately.

6. Constraining the search space

A number of authors have pointed out the difficulty of postulating an unconstrained distributional learning model, because of the enormous number of surface correlations that would need to be computed [Pinker (1987), Maratsos (1998, p. 447), Bloom (1999, p. 287)]. Gordon (1988) calculated that a child would have to sift through 8 billion possible contexts to learn the mass/count distinction, yet children command mass/count syntax by age 2;6. What are the constraints on the learning system that allow children to rapidly arrive at appropriate linguistic categories? What allows children to home in so quickly on the relevant cues?
Among the proposed solutions to this problem is, of course, the idea that children come equipped with innate linguistic knowledge, with a “restricted property register” [Maratsos (1998)] that considers only certain properties as grammatically relevant. One might also think of innate processing constraints. Children might be geared towards looking for some types of contexts and not others. For example, a child would register whether a determiner agrees with the noun, but will not try to compute whether the noun agrees with the last morpheme in the sentence [Pinker (1987)]. A similar effect is obtained if children restrict their search to phonological phrases [Peters (1997)].

Aslin, Saffran and Newport (1999, p. 378) propose to replace innate constraints by innately biased statistical learning mechanisms but they do not specify what these innate biases are. Other authors focus on the learner’s attention span. Finch and Chater (1993) suggest that children consider only adjacency relations; Gerken and McIntosh (1993) propose that they focus only on the location of words relative to frequent morphemes [see also Valian and Coulson (1988)]. In the same vein is Elman’s (1993) work showing that a connectionist model succeeds better if its initial input is limited. Addressing this problem experimentally, Santelmann and Jusczyk (1998) started to explore children’s capacity to identify dependencies between elements such as the auxiliary is and the ending -ing on the verb (the dog is running but not *the dog can running). They find that 18-month-olds, but not 15-month-olds, have developed a sensitivity to this dependency. They also show that the capacity to track the dependency remained when a two-syllable adverb intervened between the auxiliary and the verb (John is always running), but not when the adverb had three or four syllables. This suggests that children initially register only adjacency-type relations; as their attention and short-term memory span develop, they become able to notice non-adjacent dependencies and to track them across more and more distant positions.

The correct solution is probably a combination of these factors, and possibly still others.

7. Conclusion

To sum up, it is clear that language input orients children toward discovering the semantic and syntactic categories coded by the language they are learning. Children exploit every possible source of information at their disposal to determine the meaning and the grammatical category of a word. They pay attention to formal cues from the beginning. The cues they initially use are probably partial and tentative. Children probably first notice statistical regularities and use them as cues in parsing language, without necessarily having a sufficient mastery of these cues (or confidence in them) to be able to use them in their productive language. They may begin by noticing some co-occurrence between adjacent elements, or between some linguistic element and some aspect of the situation, and use this crude cue to help uncover other potential cues. The strength of each cue varies according to its ability to help the child parse successfully. Gradually, children become more able to track dependencies over longer stretches of discourse.
Children make few category assignment errors. They are conservative in learning the grammar of their language; they tend to use relational words in the constructions in which they have heard them used and are reluctant to generalize. Generalizations seem to appear at a second stage, when the similar behavior of members of a class is registered.

Many questions remain to be answered: What types of innate constraints are built into the system? Are some cues more salient to children than others? Just how much knowledge do very young children have of grammatical categories? How is this knowledge built up? Gordon (1988) stated that there are two points of interest in studying the acquisition of any distinction: the point of first awareness and the point of complete control. In fact, many other points of interest can be identified. Take the grammatical category of determiners. We can identify the point of first awareness of determiners, the point where determiners start to be used in parsing the clause, the point of the emergence of fillers in production, and the point when children systematically supply determiners to introduce noun phrases. This point corresponds to adequate basic use, but Karmiloff-Smith (1979) has shown that correct use is not the end of learning, and that it takes many years before determiners are integrated into a system of relevant markers and are used proficiently in discourse, taking the pragmatic context into account. Language acquisition studies of grammatical categories face many years of intensive work.

References


Ch. 19: The Acquisition of Grammatical Categories


Clark, E. (2005), (this volume)


Ch. 19: The Acquisition of Grammatical Categories 453


Chapter 20

SEMANTIC CATEGORIES IN ACQUISITION

EVE VIVIENNE CLARK

Stanford University

Contents

Abstract 459
1. Introduction 460
2. Space 461
3. Shape 465
4. Adding common ground 466
5. Conceptual domains and lexical options 467
6. Adding meaning in the course of conversation 472
7. Universals in mapping? 473
8. Conclusion 476
References 477

Abstract

Both universal and language-specific meanings play a role as children map semantic categories onto linguistic forms in a first language. What sources do they draw from as they do this mapping? To what extent are their semantic categories informed by universal conceptual categories or by the conventions of the language community? In this chapter, I consider some of the contributions of cognitive and social factors in children’s construction of semantic categories.
1. Introduction

How children map language onto conceptual structure depends on at least two factors: the conceptual categories they bring to the task – in the form of knowledge they have already acquired – and the adult uses of the target language. Children depend on child-directed speech to discover the conventions of the language and find out which words and constructions are appropriate to convey particular intentions.

We need to first examine the extent to which there is a direct match between a priori conceptual categories and semantic categories that are “given” in child-directed speech and compare this with the extent to which children have to first learn where the boundaries of conventional terms lie relative to conceptual boundaries. In some cases, the lexical meanings in a language may map fairly directly onto the conceptual categories children have. In other cases, the mapping may be only partial, or even require the construction of a new conceptual category. While the conceptual categories that infants build up during their first year of life may be universal, the meanings expressed in each language are not. Meaning boundaries also do not always coincide from one language to another. In summary, child-directed speech plays a critical role in the construction of lexical categories during acquisition. And what adults say to children in conversation offers them a major source of information about the semantic categories of the language they are learning.

Are lexicosemantic categories mapped directly onto preexisting conceptual categories? This would seem reasonable, in the case of children’s early word meanings, but as they learn more language, they will also encounter words that denote categories they have not yet begun to represent. Therefore, some categories have to be instead built from the words they hear. Nonlinguistic experience does not provide all the categories a priori that are represented in the conventional lexicon of a language. For example, verbs for actions, including any associated objects, often involve culture-specific linguistic constructions, so children must then construct the relevant concepts by observing just how the relevant words are used in adult speech. Of course, children could simply relate every unfamiliar word meaning to the closest candidate category they already have represented, only later extending the word meaning to match adult usage.

Because languages differ, the boundaries of conceptual categories and semantic categories can diverge at many points. Consider the differences in the numbers of terms people use for talking about containers [Malt et al. (1999)]. When English, Chinese, and Spanish speakers were each shown 60 household containers for liquid, the three groups came up with different numbers of labels, as shown in (1)–(3). The boundaries for the terms speakers produced did not coincide to any great degree across the three languages.

(1) English speakers – 7 terms
   — jar (19), bottle (16), container (15), can (5), jug (3), tube (1), box (1)

(2) Spanish speakers—17 terms
   — frasco/frasquito (28), envase (6), bidon (6), aerosol (3), botella (3),
   — pote/potecito (2), lata (2), taro (2), mamadera (2), caja (1), gotero (1),
   — pomo (1), roceador (1), talquera (1), taper (1)
(3) Chinese speakers – 5 terms
– ping2 (40), guan4 (10), tong3 (5), he2 (4), guan3 (1)

Yet, when sorting the same set of containers in terms of physical similarity to each other, speakers of these three languages showed much greater agreement. Their groupings of the containers revealed the extent to which perception-based categories can diverge from lexical categories. The semantic categories conveyed by word meanings do not all map in the same way onto the conceptual categories in the domain of containers. And since children and adults can sort, group, and categorize in many contexts, with and without language, they must rely on multiple representations for objects, properties, relations, actions, and events, some driven by spoken (or written) language, and some independent of language [Gentner and Goldin-Meadow (2003)]. But people draw on those representations that are relevant to language whenever they wish to talk (or write) about what happened, and what they saw, remembered, or did [Slobin (1996)].

2. Space

Before children understand the word meanings relevant to a specific conceptual domain, they can draw only on what they already know about the entities involved. Let us look at, for example, the concept of space and the conceptual representation of spatial relations [Clark (2004a)]. When do infants discriminate different conceptual relations in space? Recent studies show that for some spatial relations, they do this quite early. Six-month-old infants readily habituate to pictures of a container from several angles and holding different objects, so they presumably accept the container in some sense as exemplifying the same relation on each occasion [Casasola, Cohen and Chiarello (2003)]. A few months later, infants discriminate other spatial relations too. For example, between age 9 and 14 months, in addition to a notion of containment, infants come to discriminate between tight-fit containment (e.g., a tape in a cassette case) versus loose support (an apple on a plate), and also between tight-fit containment versus loose containment (an apple in a bowl). This suggests that by around age 12 months, infants are sensitive to a variety of conceptual spatial contrasts, yet only some of these will end up being relevant to the language they will learn [Choi et al. (1999), Casasola and Cohen (2002), McDonough, Choi and Mandler (2003)].

The discriminations infants can make correspond quite closely to how they react in spatial tasks designed to assess their knowledge about space [Clark (1973b, 1980)]. Consider how 15- to 18-month-old infants dealt with the spatial relations encoded by the English prepositions in, on, and under. When an adult asked infants this age to put a toy mouse in or on a box, the infants always placed the mouse in the box, regardless of how the adult worded the instruction [Clark (1973b)]. If the adult presented them with a box lying on its side, and again asked for a small toy to be placed either in or on the box, both 1- and 2-year-old children first turned the box so its opening faced upward. Asking children to put the mouse on this box elicited the same response: they
turned the box so its opening faced upward and then put the mouse inside. The consistency of these actions strongly suggests that young children attend to the canonical orientation of objects (here, the box) and placement (the toy mouse inside), but they do not yet understand either in or on.

When the same children were given further tasks where the adult asked them to copy just what she does, they had no difficulty reproducing an array where a toy mouse was placed in an upright glass, but they “failed” when it was set on top of an upside-down glass, or next to an upright glass, or when the mouse was placed so that there was a small gap (2 cm) between it and the glass: they always placed the mouse inside the upright glass [Clark (1973b)]. They had similar difficulties when the mouse was placed 2 cm away from a block – they placed it on top of the block. In short, 1- and 2-year-olds rely first on their conceptual knowledge in responding with plausible customary spatial relations between objects and locations.

This knowledge appears to derive from infants’ experiences of how objects and places are usually connected. They know that small objects go in containers or on supporting surfaces, that objects normally touch a target location rather than stand apart from it, and that objects normally appear in their canonical orientation. Boxes, glasses, and cups stand with their openings facing upward; tables, shelves, and other supporting surfaces are horizontal; objects with an inherently vertical orientation stand with their top uppermost, and objects (e.g., cars, or trains) that can move when they are driven, generally move forwards (e.g., Ghent (1961), Rock (1973), Kuczaj and Maratsos (1975), Clark (1980), Levine and Carey (1982), Feist and Gentner (1998), Hespos and Baillargeon (2001), Meints et al. (2002), see also Gentner and Feist (in press). Young children also treat target locations as goals of motion rather than as sources of motion [Pléh (1998)]. By the age of 12 months, infants know about many kinds of relations from which to generalize about spatial relations connecting objects and places. They know, for example, that metal caps go on bottles, and bottles stand upright; that cushions go on beds or chairs, shoes go on feet, and hats go on the head; that boats float on water, cars drive on roads, and trains go on rails; and that toys go in boxes or bags, that bags are carried by the handles, and so on. In fact, much of what 1- and 2-year-olds know about physical space can be assessed from their strategies for placing objects in space. Some of this information is summarized in Table 1.

As children start to learn the terms for specific spatial relations in their language, treatment of spatial relations may start to diverge from one language to the next. Consider the preposition on in English. Cups can be on the table, or someone can put them on the table: tables are prototypical supporting surfaces. But in addition, pictures can hang on walls, cups and doors have handles on them, geckos can sit on the ceiling and flies can land on a lizard. On also has non-spatial uses, e.g., on Monday, on the 21st, on time; on a whim, on impulse, on cue, on pitch, on course, and on the sea. In short, in English, on is used for more than just support on a horizontal surface. The same is true for in, e.g., in May, in 1979, in time; in a rush, in haste, in spate; in accord; in fear, in anticipation, in a rage, in a state; as well as in the box, and in the room. The same is also true for other spatial prepositions.
Other languages make the situation even more complicated. Consider the three relations pictured in Figure 1: support—a cup sitting on a table; containment—an apple in a bowl, and attachment—a handle on a cupboard door. The English language uses *on* for the relations of support and attachment (cups on tables and handles on doors), and *in* for containment (apples in bowls). Spanish uses just one preposition, *en*, for all three relations: support, attachment, and containment. Dutch uses three prepositions: *op* for support, *aan* for attachment, and *in* for containment. And Finnish, a case-marking language, uses two case forms on the noun for the location: the inessive case ending –*ssa* for attachment and containment, and the adessive case –*lla* for support [Bowerman (1996)]. Therefore, children with various native languages have to learn what their languages’ semantic categories are, compared to any conceptual categories they may have set up in their minds on the basis of nonlinguistic information and experience. Although infants of various native languages probably start out with much the same conceptual representations for space in their first year, they will often follow different paths as they learn how their particular language maps onto the conceptual domain of space.

Let us look at another example. Imagine a coat hanging *on* or from a hook, a mobile hanging from the ceiling, and two toy train wagons linked together. These three relations are described in English using combinations of verbs and prepositions or adverbs: coats and mobiles *hang* *on/from* a hook, and from the ceiling, respectively, while the wagons for toy trains *hook* *together*. In Korean, speakers divide things up differently. They use one verb, *kelta*, for hanging a coat on a hook and *talta* for attaching the train wagons, and another verb, *talta*, for the mobile suspended from the ceiling.

When children learn to talk about relations like these, they follow the patterns adults favor for actions of placing and joining [Bowerman, de León and Choi (1995)]. Adult speakers of English typically rely on a general-purpose verb (usually *put*) combined with the preposition *on*, and rely less often on verbs that combine with *in* (usually *put*

<table>
<thead>
<tr>
<th>Spatial Knowledge: Early discriminations and strategies for coping with spatial relations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Early discriminations among spatial relations</strong></td>
</tr>
<tr>
<td>Containment: from age 6 to 7 months</td>
</tr>
<tr>
<td>Loose containment: from age 9 to 14 months</td>
</tr>
<tr>
<td>Tight-fit containment: from age 9 to 14 months</td>
</tr>
<tr>
<td>Loose support: from age 9 to 14 months</td>
</tr>
<tr>
<td>Support: from age 17 to 19 months</td>
</tr>
<tr>
<td><strong>Strategies used starting at age 15–18 months to relate an object “X” to a location</strong></td>
</tr>
<tr>
<td>Containment: If the location is a container, put X inside.</td>
</tr>
<tr>
<td>Support: If the location is a supporting surface, put X on top.</td>
</tr>
<tr>
<td>Contact: If the location is nearby, put X in contact with it.</td>
</tr>
<tr>
<td>Movement goal: If the location is mentioned, move X toward it.</td>
</tr>
<tr>
<td>Canonical orientation: If the location is not canonically positioned, return it to its normal orientation and so on…</td>
</tr>
</tbody>
</table>

Table 1
again) or *together* (also *put*, but sometimes *hook* or *fasten*). For the same domain – actions of placing and joining Korean-speaking adults use some 16 distinct verbs, with the choice of verb determined by the object being placed. [Like Japanese, Korean relies on different verbs for the donning of clothing, depending on which body part is affected; see Backhouse (1981)]. In the English and Korean languages, the lines distinguishing one verb from another only really coincide for the English verb *close* and the

---

Fig. 1. Support, containment, and attachment in four languages (based on Bowerman 1996: 394).
Korean verb *tatta*. By age 2 years and 6 months to age 3, children acquiring Korean are well on their way to acquiring several Korean distinctions, with their two main verbs (both somewhat overextended) being *kkita* (‘fit together, mesh with’) and *nehta* (‘put loosely in/around’). The children already use a few specialized clothing verbs, including *sinta* (‘put … on feet’) and *ssuta* (‘put … on head’) [see Kameyama (1983), Bowerman and Choi (2001) for the acquisition of Japanese verbs of clothing].

Even when two languages are closely related, speakers do not necessarily do things in the same way linguistically. Bowerman compared how English and Dutch speakers talk about the action of separating two objects. In English, adult speakers use *out* for removing something from a container, and *off* for removing clothes. This distinction is well established even in young 2-year-olds. In Dutch, adults also rely on two particles, *uit* and *af*, but their assignment differs from that of the English *out* and *off*. *Uit* is used for removal of clothing, whereas English assigns *off* for this meaning. Dutch 2-year-olds overextend *uit* to cover the domains of both *uit* and *af* [see further Bowerman et al. (1995), Bowerman and Choi (2001)].

### 3. Shape

Just as children rely on what they already know about spatial relations, using their conceptual representations of objects and places, they also rely on what they already know when identifying, sorting, or grouping objects. And in trying to talk about things for which they do not yet have a word, they look for similarities in the shape of things. This is most in evidence as children begin to produce their first words. They overextend early words, and, for example, use a term like *dog* not only for dogs but also for other four-legged, mammal-shaped entities as well, or a term like *stick* for sticks and for rulers, canes, furled umbrellas, and other long thin things [Clark (1973a), Anglin (1993)]. They also make use (though to a lesser extent) of properties like size, sound, characteristic motion, taste, and texture. Some typical examples of these early overextensions are shown in Table 2.

The preference for shape children display in their overextensions between 1 year, 8 months to 2 years, 6 months also shows up in sorting and categorization tasks when young children have to match one of two objects to a sample. Shape consistently takes priority over dimensions like color in both younger and older preschool children [e.g., Landau, Smith and Jones (1988), Baldwin (1992), Imai, Gentner and Uchida (1994)]. As children get older, they also start making use of other kinds of information. In particular, they appear eager to make sense of where things fit, and start to solicit information about function between age 3 and 4, with questions like “What’s that for?” [Kemler Nelson, Egan and Holt (2004)].

In summary, children attend to shape early on and discover that it is particularly useful for object categorization [Landau et al. (1988)]. They can then make use of shape as they assign initial meanings to the words they are learning to map to their conceptual categories.
4. Adding common ground

Just how do children start on the mapping they must do if they are to learn how to talk about the world around them? To identify probable meanings, they need to be able to establish common ground with the adults talking to them. In adult–adult exchanges, the essential conditions speakers and addressees observe in establishing and then accumulating common ground include joint attention, physical copresence, and conversational copresence [Clark (1996)]. With young children, adults work to establish joint attention, often using several rounds of name-use, other verbal attention-getters, and gestures to get the child’s attention focused on what the adult plans to talk about [Estigarribia and Clark (in preparation)].

Adults commonly take as the topic of the conversation objects and events that are physically accessible – physical copresence – and hence conversationally copresent as well. Attention on the child’s part is integral to joint attention and to any subsequent uptake of linguistic material as the child and adult accumulate common ground [Tomasello (1995), see also Clark (2001b, 2004b)]. In fact, children pay close attention to what is physically present when they learn a new word [Aslin and Smith (1988)].

Table 2
Overextensions based on shape, size, motion, sound, and texture [based on Clark (1973a)]

<table>
<thead>
<tr>
<th>Word</th>
<th>First referent</th>
<th>Domain of (over)extensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>mooi</td>
<td>moon (Eng.)</td>
<td>&gt; cakes &gt; round marks on windows &gt; writing on windows and in books &gt; round shapes in books &gt; tooling on leather book covers &gt; round postmarks &gt; letter O</td>
</tr>
<tr>
<td>nénin</td>
<td>breast (Fr.)</td>
<td>&gt; button on garment &gt; point of bare elbow &gt; eye in portrait &gt; face in portrait &gt; face in photo</td>
</tr>
<tr>
<td>gumene</td>
<td>coat button (Serb.)</td>
<td>&gt; collar stud &gt; door handle &gt; light switch &gt; any thing small and round</td>
</tr>
<tr>
<td>baw</td>
<td>ball (Eng.)</td>
<td>&gt; apples &gt; grapes &gt; eggs &gt; squash &gt; bell clapper &gt; anything round</td>
</tr>
<tr>
<td>kottiebaiz</td>
<td>bars of crib</td>
<td>&gt; large toy abacus &gt; toast rack &gt; picture of columned façade</td>
</tr>
<tr>
<td>tee</td>
<td>stick (Eng.)</td>
<td>&gt; cane &gt; umbrella &gt; ruler &gt; (old-fashioned) razor &gt; board of wood &gt; all sticklike objects</td>
</tr>
<tr>
<td>mum</td>
<td>horse (Eng.) toy</td>
<td>&gt; cow &gt; calf &gt; pig &gt; moose &gt; all 4–legged animals</td>
</tr>
<tr>
<td>ass</td>
<td>goat on wheels,</td>
<td>&gt; sister &gt; wagon (things that move) &gt; all things that move &gt; all things with rough surface</td>
</tr>
<tr>
<td></td>
<td>with rough hide (Ger.)</td>
<td></td>
</tr>
<tr>
<td>fly</td>
<td>fly (Eng.)</td>
<td>&gt; specks of dirt &gt; dust &gt; all small insects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt; child’s own toes &gt; crumbs of bread &gt; toad</td>
</tr>
<tr>
<td>fafer</td>
<td>chemin de fer (Fr.)</td>
<td>&gt; steaming coffeepot &gt; anything that hissed</td>
</tr>
<tr>
<td></td>
<td>sound of train</td>
<td>or made a noise</td>
</tr>
<tr>
<td>wau-wau</td>
<td>dog (Serb.)</td>
<td>&gt; all animals &gt; toy dog &gt; soft house slippers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt; picture of old man in furs</td>
</tr>
</tbody>
</table>
Adults co-opt infants into dyadic interactions early, incorporating them in protoconversations long before they can produce any words. Instead, adults count kicks, burps, smiles, and babbles as infant-turns. Once infants begin to offer consistent babbles and even words as turns, adults will accept only those. Burps and kicks no longer count [Snow (1977)], as shown in (4):

(4) Mother and Ann (aged 1 year, 6 months)
   Ann: (Makes blowing noises.)
   Mother: That’s a bit rude.
   Ann: Mouth.
   Mother: Mouth. That’s right.
   Ann: Face.
   Mother: Face. Yes. Mouth is in your face. [Snow (1977), p. 18]

In setting up speaking turns for small children, adults work simultaneously on achieving both joint attention, hence a joint locus of attention, and physical and conversational copresence in their interactions. Infants may only really begin to participate in turn-like speaking alternations at around 10 months of age, and begin to make their turns contribute to the interaction only from 12- to 18-months of age and older. It is in conversational contexts that adults introduce words for objects, actions, and relations as they talk to small children about events and states, and the participants in those events – and thereby display some conventional ways of talking about such things [e.g., Clark (1999, 2003a)]. Once adults can establish these words as part of common ground, children can begin to work on discovering the conventional meanings of the words adults offer. This route to conventional meanings is critical for the acquisition of the lexicon in any language.

5. Conceptual domains and lexical options

For children learning the conventional lexicon of a language, a primary problem is the mapping of conceptual categories and linguistic forms. They must work out what each form refers to, and how the meanings of neighboring forms are related to each other. Child-directed speech is critical: children attend to and follow adult usage [e.g., Bowerman (1985), Bowerman and Choi (2001)]. This enables them to group terms for objects and events that belong in the same conceptual domain – terms for either animals, foods, or vehicles, for example, or for either caused motion, inherent states, or changes of state.

Adult usage guides children in mapping the boundaries of domains and of the constituent categories in those domains. In effect, the boundaries given by the conventions of each language separate out the semantic ‘bundles’ represented by each word in the lexicon of that language. At the same time, the way that speakers group objects on the basis of physical similarities differs from how they group them on the basis of the words they use [e.g., Swanenflugel and Rey (1986), Malt et al. (1999)]. In many domains, the words that speakers use have different histories in each language; the number of words
in the domains differ, along with the actual kinds of entities (e.g., birds, vehicles, or containers) that are present. Therefore, for speakers, the available word choices differ too.

One also finds subtly different histories for structural elements in languages. To take one example, consider how auxiliary have and be are assigned for talking about event types, where these appear to form a continuum, as shown in Table 3. Note that change of location, with the auxiliary be, is at one end of the continuum, and controlled (non-motion) process, with the auxiliary have, is at the other end of the continuum, both acting as anchors; there is little variation in these anchors across languages. But the changeover point from the auxiliary be to the auxiliary have varies from one language to the next, even when the languages are closely related [Sorace (1993, 2000)].

To reiterate, semantic categories do not map to each other across languages in any direct fashion. Any crosslinguistic mapping depends on the number of terms in the target domain, where the domain boundaries lie, and the cultural history of the language as this is reflected in borrowings, changes of meaning, size of contiguous domains, and so on.

The boundaries for lexicosemantic categories are likewise seldom clear-cut as one moves from one language to another. Even within languages, they can shift from one task to another. In one study, adults were shown line drawings of an array of 19 cup-like objects, all with a single handle on the side, and varying systematically along several dimensions: diameter of the bowl of the cup relative to its height, whether its sides curved in toward the base or went straight down, whether it had a foot or even a stem, and whether the container was round or either 3- or 4-sided. Each picture was presented singly, in either a neutral context or a food context. In neutral contexts, adults preferred to use cup when the height and width were either in a 1:1 or 1:1.5 ratio, but shifted to bowl as the width increased relative to height. In food contexts, speakers’ choices of cup vs. bowl depended on a combination of the verb (eat vs. drink), plus the noun for any implement mentioned (spoon), and whether the food-type was liquid or solid [Labov (1973)]. But this study looked only at line drawings rather than real containers, and did not take into account the effects of material.

<table>
<thead>
<tr>
<th>Event type</th>
<th>Auxiliary choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change of location</td>
<td>BE (least variation)</td>
</tr>
<tr>
<td>Change of state</td>
<td>↑↑</td>
</tr>
<tr>
<td>Continuation of preexisting state</td>
<td>↑↓</td>
</tr>
<tr>
<td>Existence of state</td>
<td>↑↓</td>
</tr>
<tr>
<td>Uncontrolled process</td>
<td>↑↓</td>
</tr>
<tr>
<td>Controlled process (motion)</td>
<td>↓↓</td>
</tr>
<tr>
<td>Controlled process (nonmotion)</td>
<td>HAVE (least variation)</td>
</tr>
</tbody>
</table>

Table 3
Auxiliary selection hierarchy
With boundaries, children must learn what the conventions are, for their language, for the use of each expression in the target domain as well as the factors that affect their choices. This takes time. For instance, how do children assign the terms *cup* and *glass* in English? Andersen (1975) explored this question by asking children aged 3, 6, 9, and 12 to name and then to sort 25 different drinking vessels (real vessels, not pictures). All the terms produced at each age for the objects, are shown in Table 4. At the end of the experiment, she also asked the children to give definitions for *cup* and *glass*, and to pick out the best exemplar for each from the array in front of them.

The children appeared to go through three stages in learning to assign *cup* and *glass* appropriately. In the first stage, at age 3, they overextended the word *cup*, using it for many of the drinking vessels without regard to proportions (height to width ratios), material (ceramic, glass, or plastic, etc.), the presence of a handle (e.g., cups vs. glasses), or the customary contents (e.g., coffee vs. milk vs. wine). In the second stage, at age 6, they focused primarily on perceptual properties and used only those properties in deciding whether a container was a cup or not. Then, in the third stage, at age 9, they started appealing to functional properties as well, taking into account what the container is generally used for (e.g., milk, water, coffee, wine, etc.). At this point, the children also showed some awareness that the boundary separating cups and glasses is a fuzzy one. For example, 9-year-olds commented on the fact that glasses do not have to be made of glass, and paper cups need not have handles.

But when the same children were asked to sort the 25 objects, instead of simply to label them, they shifted the boundaries for the groupings made with the labels. An item that had been labeled *cup* was sorted with items labeled *glass* between 12% and 22% of the time, and an item labeled *glass* was sorted with items that had been labeled *cup* between 3% and 7% of the time. With age, the number of such shifts declined, and began to depend less on the material the container was made of. But, not surprisingly, the boundaries for sorting often did not match those for labeling [Andersen (1975), see also Malt et al. (1999)].

Lexical boundaries appear even harder to discern for categories of actions than for categories of objects [Gentner (1982)]. Take the case of the verb *open*. Its meaning has to be constructed from the objects it is applied to – and, in English, *open* applies to windows, doors, boxes, jars, bottles, briefcases, cupboards, drawers, desks, ovens, and so

### Table 4
Terms for drinking vessels used at each age

<table>
<thead>
<tr>
<th>Age</th>
<th>Terms produced and number of children using each term</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-year olds (n = 6)</td>
<td>cup – 6, glass – 3, mug – 1, bowl – 1, a pepper – 1, ashtray – 1</td>
</tr>
<tr>
<td>6-year olds (n = 6)</td>
<td>cup – 6, glass – 5, mug – 4, dish – 3, bowl – 1, vase – 1, pitcher – 1, sugar – thing – 1, hourglass – 1, funnel – 1</td>
</tr>
<tr>
<td>9-year olds (n = 6)</td>
<td>cup – 6, glass – 6, mug – 6, dish – 2, bowl – 2, vase – 1, pitcher – 1, thingamajig – 1, a measure – 3, can – 1, jigger – 1</td>
</tr>
<tr>
<td>12-year olds (n = 5)</td>
<td>cup – 5, glass – 5, mug – 5, dish – 1, bowl – 1, vase – 1, wine-sipping thing – 1, a measure – 1, egg holder – 1, pencil holder – 1</td>
</tr>
</tbody>
</table>
on. How do young children use this verb? Some typical examples are shown in Table 5. The first use listed appears quite conventional, but the second and third reveal that the child’s meaning includes a general notion like “remove obstruction,” or “facilitate access to.” The challenge is to work out where to use open and where to rely on another verb (move out, tilt up, bite, extend, pull up, peel, get out) for actions performed on particular objects [Clark, Carpenter and Deutch (1995), Goodman, McDonough and Brown (1998), Wilkins (2002)].

The boundaries of the meanings of words for actions may have to be derived by children from their knowledge about the kinds of objects affected. For example, where adults used a verb for an action that applies to a relatively small coherent set of objects, children seem to have little difficulty grasping the dimensions that are relevant to the action, and so arriving at the adult semantics for that verb. They make few errors in applying the verb themselves, and so master the mapping of word to action fairly easily. But when the word for an action applies to a large, disparate set of objects from which it is more difficult to derive the precise dimensions relevant to the action, children seem to have a much harder time learning how to use the pertinent verb. The semantics for that verb does not map onto any neat a priori conceptual category. This is what appears to cause young children difficulty in the acquisition of English verbs like open, cut, and break [Bowerman (2005)].

Languages differ in what is lexicalized in a verb. Take the case of verbs for motion events. These can contain information about manner and information about path. Compare the English verbs stroll, meander, trot (motion/manner) and enter, exit, raise (motion/path), with walk into the house, climb the tree, run out along the rocks (all motion/manner/path). For the most part, English tends to combine motion and manner into a single lexical package, with information about path given in a prepositional phrase. The Romance languages tend to combine motion and path, and leave information about manner for an adverbial phrase, a separate clause, or simply to be inferred from other details [Slobin (2001), Gennari et al. (2002)]. In learning the conventional meanings of verbs, children must work out which properties of motion events are packed into the verb, and which properties must be expressed some other way.

Table 5
Early uses of the verb open by a young boy, “D,” from age 1 year, 9 months to age 2 years, 4 months (1;9.1 – 2;4.2)

- trying to get the lip of a milk carton open: I open.
- trying to squeeze past the corner of a table: Open!
- wanting a chair pulled out from the table: Open a chair.
- tilting a bowl so that only part of the bottom rested on the tabletop: D open a bowl.
- taking a bite out of his toast: Get a open it!
- having extended his fingers: I open a finger.
- pulling up his T-shirt to display his stomach: I open my T-shirt.
- wanting his father to peel a tangerine: Herb, open that tangerine. Open that.
- trying to get his napkin out of its ring, says to his mother: You open this?
In studies to determine which differences infants can detect, Pulverman and her colleagues (2003, 2004) found that 7- to 8-month-old infants detected changes in path but not in manner, when viewing film clips of motion events. But 14- to 17-month-olds detected not only changes in path but also changes in manner and in combined path+manner. This suggests that young children attend to the relevant dimensions in motion events. This is a prerequisite for mapping verb meanings for those events. They need to be able to identify the properties encoded by each verb; they must also identify any consistent patterns of conflation for motion+manner or motion+path in the meanings of specific verbs, and they must be alert to any other ways to express manner and path.

Languages differ in just which dimensions of meaning are to be clustered or bundled in a single lexical item. Children learning different languages exhibit different degrees of difficulty in learning verb meanings in certain domains, because they have to identify just which dimensions are pertinent. One way they may have to do this is by keeping track of which objects can be affected by an action. This in turns requires keeping track of the set of objects mentioned in conjunctions with a specific verb in conventional adult usage, in that language [Bowerman (2005), see also Marcotte (in press)]. The semantic category comprised by the verb does not always map very well onto the kinds of conceptual action categories already identified. Adult usage shapes offers children semantic categories of actions and of objects that can differ from language to language. They also, typically differ in many details from whatever conceptual categories children may already have set up.

To summarize, children learn the conventions for the language around them. Both the words available and the patterns of use in child-directed speech play a critical role here in shaping children’s lexical categories. While children may often start from conceptual categories that have already been formed, newly-introduced lexical categories may themselves also be precursors to other conceptual categories. That is, development may go in both directions, from conceptual categories to word meanings and from word meanings to conceptual categories. This in turn strongly suggests that children retain distinct representations – conceptual versus lexical, for example – that they can put to use under different circumstances. Consider sorting tasks where we need to put together objects that are similar, versus labeling tasks where we need to distinguish each type in a set from its neighbors, versus acceding to requests on occasions where we need to identify the intended referent of a speaker’s utterance. Each task draws on a different type of representation [Gentner and Goldin-Meadow (2003)]. Essentially, children make use of what adults offer – about what things are called, how they are related to other objects, where they stand in space, how they move, and so on.

In doing all this, children observe many of the same conditions on conversational exchanges that adults do, as they master a growing number of the conventions specific to their particular language. They attend to pragmatic directions about language from adults, and they take up adult offers both of words and of information about the referents of those words. So the setting for acquisition plays a critical role in the acquisition of both conceptual and lexical categories.
6. Adding meaning in the course of conversation

Consider the conditions that govern conversational exchanges: speaker and addressee have joint attention; they may also have physical copresence (the locus of their joint attention), and they have conversational copresence. These conditions allow speaker and addressee to achieve common ground and to add to common ground with each piece of new information contributed as the exchange proceeds [Clark (1996)].

Young children take up adult offers of words, for example, and show they are doing so by repeating the word in question, as shown in (5) and (6) [Clark (2002, 2004a)]:

(5) Mother (looking with a child, at a picture of some owls):
What are these? Those are birdies.
Child (Age 1;7.19): Birdies.

(6) Sarah [(Age 3;6.6), looking a picture of a nest with eggs in it]
Mother: That’s a nest.
Sarah: A nest.

Along with new words, children also receive information about the referents of the words, as shown in the underlined words in (7) and (8) [Clark (1999), Clark and Wong (2002)]:

(7) Sarah [(Age 3;6.6), looking a picture of a nest with eggs in it]
Mother: That’s a nest.
Sarah: A nest.
Mother: Um. That’s where the birdies live. That’s a birdie house. They call it a nest.

(8) D [(Age 3;9.18), at the airport; watching as a mechanic puts two chocks by the plane wheels]: Why did he put two loggers?
Mother: Oh, they’re called “chocks,” and they keep the wheels from moving.
D: Why did he put the chocks? [Clark (unpublished diary data)]

Notice that in making offers of words, adults are presenting children with the conventional term for the entity in question, and so underlying the assumption of conventionality for all speakers in that community, namely that “for certain meanings, speakers assume that there is a conventional form that should be used in the language community” [Clark (1993), p. 67]. This assumption goes hand in hand with contrast, where both speakers and addressees “assume that any difference in form signals a difference in meaning” [Clark (1993), p. 69].

These assumptions also come into play when children make mistakes not just in word choice but in also in morphology and syntax, as in the side sequence in (9) and in the embedded correction in (10) [Chouinard and Clark (2003)]:

(9) Abe (Age 2;5.7): The plant didn’t cried.
|| Father: The plant cried?
|| Abe: No.
Father: Oh. The plant didn’t cry.
Abe: uh huh.
(10) Abe (Age 2;5.10): *I want butter mine.*
   Father: OK. Give it here and I’ll put butter on it.
   Abe: *I need butter on it.*

Indeed, children make use of this information that adults offer when simply talking to them and when checking up on what their children had intended to say.

7. **Universals in mapping?**

To what extent do children build on conceptual universals that then become mirrored in lexical universals? [see Greenberg (1966), Clark and Clark (1978), Bybee (1985), Bybee, Perkins and Paglinca (1994)]. Even when children receive no lexical expression in some languages, certain conceptual distinctions appear to be so salient that they try to map them onto linguistic forms and so may construct temporary mismatches with the conventions of the language. One way to approach this is to distinguish robust categories and emergent categories for any one language, and then to take the conjunction of the two categories as marking the candidates that are the most salient, or most general, or even the best ones for universal conceptual categories.

Robust categories are categories that turn up in a specific language and are fully supported by child-directed speech. Many robust categories are universal in that they are strongly represented across languages [e.g., Bybee et al. (1994)], and appear in some form in nearly every language. But there are also “gaps” in certain languages where particular distinctions lack any conventional realization. In those languages, those distinctions sometimes surface as emergent categories in early language acquisition [Clark (2001a)].

I have called temporary distinctions drawn by young children **emergent categories**, conceptual categories that are particularly salient and in other languages often receive conventional form, typically as grammatical categories of some type. They are conceptual categories so salient that young children look for ways to express them even when child-directed speech offers no guidance. Examples of such categories include the notion of “source,” which underlies notions of agency, natural force, cause, possession, and comparison [Clark and Carpenter (1989a,b)]. Children may pick up on a single lexical expression to connect all these sources, as illustrated in Table 6. As they learn the different lexical expressions here, for each subtype of source, they relinquish their earlier reliance on a form like *from* to mark the more abstract notion of source, and replace it with the conventional forms in use in the speech community.

Another candidate emergent category in the speech of some children is the distinction between inherent and temporary properties [Heyman and Diesendruck (2002), Sera (1992)]. Some languages distinguish inherent properties from temporary ones that result from some activity that produces a change of state. While English lacks a conventional means for making this distinction, children sometimes impose it, recruiting for that purpose a distinction in form found among adjectives. Some adjectives are formed from nouns with the suffix -y, while others are formed from verbs with -ed, as shown for the novel adjectives listed in Table 7.
In context, most of the coinages in -y appeared to designate an inherent property, while most of those in -ed designated a temporary property that had resulted from an observable action [Clark (2001a)]. Moreover, the forms in -y were produced in contexts where there was no observable prior action that could have been linked to the property in question. This strongly suggests that the two devices for forming new adjectives had been assigned as carrying meanings identifiable as “inherent” and “temporary,” respectively. Eventually,

<table>
<thead>
<tr>
<th>Source as an emergent category</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Agents and natural forces</strong></td>
</tr>
<tr>
<td>D (2;2.3, looking at pieces of sandwich he’d pushed off the edge of his plate): These fall down from me. (= I made these fall down)</td>
</tr>
<tr>
<td>J (2;2, recounting her visit to the doctor): I took my temperature from the doctor. [ = I had my temperature taken by the doctor]</td>
</tr>
<tr>
<td>Du (5;10.5, announcing his younger sister’s finger-puppet show): Another puppet show! From Helen!</td>
</tr>
<tr>
<td>D (2;11.12, looking at a fallen tree after a storm): Look at that knocked down tree from the wind. (= tree knocked down by the wind)</td>
</tr>
<tr>
<td>Sa (5;1) Now the rainbow is getting higher from the rain.</td>
</tr>
<tr>
<td><strong>Cause, possession and comparison</strong></td>
</tr>
<tr>
<td>D (2;11.12, recalling what he’d done three months earlier when his mother left him with his grandmother, while she fetched his father): When gran’ma ’ancy was here, You go fetch Herb. (pause) Then I cried a bit from you go get him.</td>
</tr>
<tr>
<td>S (2;8.7, explaining why his fire engine was stuck on the roof of his toy garage): That’s fro-- That’s from I put a thing on it.</td>
</tr>
<tr>
<td>S (3;0.13, of a picture of some people and a horse): That’s a finger from him.</td>
</tr>
<tr>
<td>D (3;7.5, assigning roles in a game): You can be a mum from two babies.</td>
</tr>
<tr>
<td>D (2;8.15, climbing into his car seat): This seat is getting too small from me.</td>
</tr>
<tr>
<td>W (3;1.15, talking about a toy rabbit): See, this ear is longer from the other ear.</td>
</tr>
<tr>
<td>D (3;7.1, comparing his height to his mother’s): … and you’re the tallest from me.</td>
</tr>
<tr>
<td><strong>Table 6</strong></td>
</tr>
<tr>
<td><strong>Agents and natural forces</strong></td>
</tr>
<tr>
<td>D (2;6.9, talking about amarettini biscuits): It isn’t crumby.? (= full of crumbs)</td>
</tr>
<tr>
<td>D (2;6.27, seeing news on TV): You wear gloves when it’s snowy time. (= when there’s snow)</td>
</tr>
<tr>
<td>D (2;7.5, being driven home in the dark): It’s very nighty. (= dark)</td>
</tr>
<tr>
<td>D (2;10.23, of the stone walls of a house in a ghost town): There’s a rocky house. (= house made of rocks)</td>
</tr>
<tr>
<td><strong>Temporary: new adjectives ending in “ed”</strong></td>
</tr>
<tr>
<td>D (2;4.28, after watching father break up shredded wheat): He was breaking them halved ... (= in half)</td>
</tr>
<tr>
<td>D (2;6.7, looking at the rack in the dishwasher): That fork is all all BUTTERED. (= covered in butter)</td>
</tr>
<tr>
<td>D (2;6.13, looking at pieces of veal mother had just covered in flour and put on the counter): These are floured. (=covered in flour)</td>
</tr>
<tr>
<td>D (2;6.30, getting down from the table): My foot is all crumbed. (= sole has crumbs on)</td>
</tr>
</tbody>
</table>

In context, most of the coinages in -y appeared to designate an inherent property, while most of those in -ed designated a temporary property that had resulted from an observable action [Clark (2001a)]. Moreover, the forms in -y were produced in contexts where there was no observable prior action that could have been linked to the property in question. This strongly suggests that the two devices for forming new adjectives had been assigned as carrying meanings identifiable as “inherent” and “temporary,” respectively. Eventually,
of course, the child making use of this emergent category had to give it up. It is not supported by the overall patterns of use for adjectival forms in English, even though it surfaces via the choice of copula in a language like Spanish.

A third candidate for an emergent (universal) category is children’s assignment of contrasting degree of control as meanings associated with first-person pronoun forms recorded in children aged 1 year, 8 months to 2 years, 8 months. Consider the utterances in (11) [Budwig (1989, 1995)], where me or my is used to indicate markedly greater control or agentivity, and I is used to indicate little or no control over the activities denoted by the verbs:

(11) My cracked the eggs.
    Me jump.
    My taked it off.
    My blew the candles out.
    I like peas.
    I like Anna.
    I no want those.

These children are trying to work out how such forms as I vs. me, or I vs. my differed in meaning when they were all used to refer to the first-person (the speaker). The emergent category here appears to be degree of control over the action mentioned in the verb. Children acquiring English, it seems, often begin by assigning two first-person pronoun forms different degrees of control. Compare the utterances in (11). In those with me or my, the child is the agent and controls the action; in those with I, the child lacks control over the action see Table 8.

Eventually, of course, children learn that in English, pronoun forms differ with their grammatical role, with I used for subjects and me for direct objects, while my eventually gets identified as a possessive adjective. Once they grasp this basis for choosing pronoun forms for each person, they relinquish the meaning of control they first assigned.

The combination of robust categories underlying early language combined with emergent categories make up the potential conceptual universals in language. What is identified as robust is typically the set of general categories marked in many different languages for which there is a conventional expression in language “X.” Emergent categories are those that lack conventional expression and are therefore absent from the adult forms of one language but frequently receive conventional expression in others.

<table>
<thead>
<tr>
<th>Pragmatic function</th>
<th>I</th>
<th>Me/my</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child in control of action</td>
<td>19</td>
<td>81</td>
</tr>
<tr>
<td>Child not in control</td>
<td>63</td>
<td>37</td>
</tr>
</tbody>
</table>

Based on Budwig (1989, p. 275).
The fact that some categories are emergent in the early stages of acquisition attests to their conceptual salience for young children, even in the absence of conventional linguistic expression. Lastly, further inferences about what is most likely universal, conceptually, can be drawn from typological patterning across languages and the consistency with which particular distinctions are marked linguistically [e.g., Greenberg (1966), Bybee (1985), Bybee et al. (1994), Pederson et al. (1998)].

8. Conclusion

In acquisition, child-directed speech is critical in helping children construct and shape their lexical semantic categories apart from their conceptual categories. Child-directed speech is the initial source of the conventions for each language, and this interacts with any conceptual structures already built as children become sensitive to the uses of each term within the language being used around them. These conventions of use teach children how to think “for speaking” [Slobin (1996, 2001)], as they set up the semantic categories of their language.

But adults and children alike make use of multiple representations for objects and events [Clark (2003b)], drawing on whatever is relevant on each occasion. The representation they call on depends on the task and the goal at hand – whether one is trying to remember something, sort things into groups for some purpose, organize things along specified lines (e.g., by shape, size, or function), or any other of the myriad goals for which people draw on the range of representations built up over time.

In learning a language, children learn particular conventions for the available lexical forms used to convey semantic categories. The conventions of each language, dialect, community, or even subgroup within a community, can differ in subtle ways. So speakers typically learn many sets of conventions and move with varying skill from one to another, depending on which community their interlocutor belongs to. As they master the conventions of their language, children make use of child-directed speech, offers of words and constructions, and adult reformulations of child errors. All this information about the conventions of the language is critical for learning the lexicosemantic categories of that language, and for learning how to map these to existing conceptual categories. But these semantic categories are not isomorphic with conceptual categories. Languages simply do not encode all that one can represent, or all the information that one can make use of.

The process of learning the semantic categories for a language depends on interchanges between child and adult. These are essential sources of information for children, who are (a) mapping words and conceptual categories, (b) identifying the boundaries of semantic domains and of individual terms in the lexicon, and (c) constructing additional conceptual categories when needed. As children do this mapping according to the conventions in the language being spoken, they also learn the modes of “thinking for speaking” in that language. To succeed in this process, children rely first on any universal elements used to form their initial conceptual categories. These constitute their initial
representations for the world around them. Then, as they start to attend to speech from the adults around them, they also start to assign meanings to words, and so set up semantic categories specific to that language. These semantic categories are shaped by the language that they are exposed to. Children continue to use perceptual and conceptual information apart from language, in order to sort, categorize, and remember. The extent to which people rely on language in such tasks depends on the precise goal on each occasion. While judgments of perceptual similarity are likely to be the same across languages in nonlinguistic tasks, judgments about category-membership are likely to differ when speakers use linguistic expressions to identify the target categories. This is because the distinctions embodied in the lexicon of each language are not identical across languages. Nor do they map exactly onto the conceptual categories available. In short, as children, we build up both conceptual and semantic categories. Then, as adults, we can draw on either category, depending on the task and goal at hand.

References


Chapter 21

EARLY SYNTACTIC CATEGORIES IN INFANTS’ LANGUAGE

RUSHEN SHI

Université du Québec à Montréal

Contents

Abstract 481
1. Introduction 482
2. The acquisition of grammatical categories and the earliest binary distinction of function words and content words 482
3. Input speech and the categorization of function words and content words 486
4. Function words and language acquisition 488
5. Conclusions 492
Acknowledgment 492
References 493

Abstract

The theory outlined in this chapter proposes that infants’ earliest syntactic categories correspond to the language-universal, superordinate binary distinction of content words vs. function words, and that these two categories can be derived by infants based on perceptual analysis of a constellation of acoustic/phonetic and phonological cues in the input. Function words are hypothesized to be universally minimized in spoken language. Furthermore, the theory predicts that this initial bifurcation facilitates or optimizes various important aspects of early language acquisition, including deriving refined grammatical categories, determining phrasal bracketing, segmenting words, and learning of word meanings. Data from cross-language studies of infant-directed speech, connectionist modeling, and infants’ discrimination of the two categories support the function-word/content-word distinction. The chapter also discusses experimental results suggesting that function words do impact several aspects of language acquisition. It is proposed that high-frequency function words are stored in the emerging lexicon of preverbal infants, these stored function words have the primary role of assisting early language analysis.
1. Introduction

This chapter focuses on the question of how infants start acquiring language at the earliest stage, and in particular, how they establish their earliest syntactic categories, and how these initial categories may then affect their subsequent analyses of the language input. In Section 2, I will discuss my team’s theoretical proposal, arguing that infants tackle the system of grammatical categories by first making a binary, language-universal distinction between function and content words based on the cues in their spoken forms and distribution. In Section 3, I will present experimental evidence that speech input to infants across languages contains a range of acoustical/phonetic, phonological, and distributional cues that support these two superordinate categories, and that infants are capable of detecting such perceptual cues in deriving these two categories. The discussion in Section 4 will center on the argument that the early bifurcation of words into function and content words may have a direct impact on language acquisition; experimental data on the processing and early representation of function words by infants, as well as their roles for other language acquisition tasks will also be discussed.

2. The acquisition of grammatical categories and the earliest binary distinction of function words and content words

One of the most important tasks facing children during language development is the acquisition of grammatical categories such as nouns, verbs, prepositions, auxiliaries, etc. These categories are the fundamental primitives with which larger syntactic units are formed, such as phrases, clauses, and sentences. All contemporary syntactic theories embody the notion of grammatical categories. The organization of words into appropriate grammatical categories and the rule-governed property of category grouping into larger syntactic units capture the productive nature of human languages. Indeed, according to the generative framework, native speakers of any language possess implicit knowledge of syntactic categories and rules. Such knowledge enables them to produce and comprehend utterances that they have never heard before. Even young children’s syntactic errors demonstrate the rule-like productivity of an evolving grammar. Learning to assign words to grammatical categories is therefore not only logically necessary, it must also begin before any evidence can appear in children’s production.

Several theoretical approaches have attempted to address the question of grammatical category acquisition in children. One type of theory focuses on the role of semantics in deriving the initial word categories [e.g., Schlesinger (1974), Bates and MacWhinney (1979), Pinker (1984)]. On this view, initial grammatical categories are semantically based; for example, nouns typically denote concrete objects, verbs denote actions, and adjectives denote properties. Depending on specific theoretical assumptions, the infant is viewed as either mapping words to innate grammatical categories through their canonical semantic properties [e.g., Pinker (1984)] – i.e., a process of instantiation – or deriving the grammatical categories by discovering the semantic-syntactic mapping through
inductional processes [e.g., Schlesinger (1974), Bates and MacWhinney (1979)]. Based on these theories, the co-occurrence patterns of grammatical categories in phrases and sentences play only a limited role compared with the importance of semantics or become important only later in the acquisition process; the prototypical semantic properties of grammatical categories provide the primary route toward the acquisition of these categories.

The semantically based theories contrast significantly with theories suggesting that grammatical categories can be bootstrapped by analyzing the distributional patterns among words [e.g., Maratsos and Chalkley (1980)]. Strictly distributional approaches are directly compatible with the concept that grammatical categories are abstract, formal categories defined in terms of their relative positions within utterances. For example, Maratsos and Chalkley (1980) showed that formal gender categories must be acquired on distributional grounds. Recent work suggests that parental speech input contains co-occurrence regularities that provide evidence of some grammatical categories, e.g., nouns and verbs [Mintz, Newport and Bever (2002)], and that the learner may derive these categories based on such regularities. Li (2003), for instance, using connectionist modeling, showed that the analysis of statistical regularities among words, such as between-word transitional probabilities, constitutes sufficient evidence to assign grammatical categories.

The theory that my team has proposed focuses on a putatively even more fundamental grammatical category distinction: the bifurcation of the two broad categories of function and content words [Morgan, Shi and Alloppenna (1996), Shi (1996), Shi, Morgan and Alloppenna (1998)]. Content words include nouns, verbs, adjectives, and adverbs. Function words are items including determiners, conjunctions, complementsizers, auxiliary elements (such as tense/aspect/number affixes), etc. The term “words” in this chapter refers broadly to word-like chunks from the perspective of the child, which must be minimally syllabic since the syllable appears to be the minimal processing unit for the infant [Bijeljac-Babic, Bertoncini and Mehler (1993), Bertoncini et al. (1995)]. There is linguistic evidence that refined grammatical categories show considerable language-specific characteristics. For example, they may differ in terms of the grammatical categories included. Some languages use prepositions (such as English and French), whereas others use postpositions (such as Quechua), and some use both (such as Chinese). Languages often differ in the mapping between concepts and grammatical categories, as well as in their category boundaries. For example, the distinction between verbs and adjectives in Mandarin is not as clear as it is in some other languages. Further, languages also differ in the richness of their morphology, with some using many bound morphemes for function items (e.g., the Bantu languages and the Amerindian languages), while others use only free-standing morphemes (e.g., most West African languages and many Asian languages). The existence of such cross-linguistic variability suggests that assigning words to their appropriate grammatical categories is by no means a trivial task.

Despite the cross-language variations in refined grammatical categories, we note the basic, language-universal binary distinction between content words and function words.
We propose that the rough sorting of words into the two superordinate, language-universal classes occurs prior to the acquisition of specific grammatical categories. In other words, the initial binary categories feed into the process of discovering refined grammatical categories, regardless of whether the process is semantically or distributionally based. As will be discussed later in this chapter, this initial binary sorting is not only logically plausible, but may also have direct consequences for the efficacy of a number of important language acquisition tasks, including the bootstrapping into early syntax.

In our model, we consider the link between phonology and syntax, and in particular, the phonetic and phonological cues to the distinction between content and function words. This position departs from the classical view that the relationship between syntactic categories and their phonological forms is an arbitrary one, and that phonology and syntax are independent parts of the grammar [Hockett (1966)]. We hypothesize that the most frequent function words should show a universal tendency to be minimized in their spoken form when compared to content words. This minimality may be a result of very frequent use [Zipf (1949)], low demand for meaning contrasts among function words since they make up a small set, or high degree of predictability since each syntactic position requiring a function word permits only a limited number of word choices. Although we consider the minimality of function words to be universal, the manifestations of this minimality are universal at some levels, but are subject to language-specific constraints at the more abstract phonological level (this will be discussed in more detail in Section 3). We further propose that the language learner can derive the basic distinction between function words and content words on the basis of the perceptual analysis of multiple phonetic, phonological, and distributional facts. We suggest that the rudimentary categories of function words and content words may assist the language learner in breaking into the syntactic system and performing tasks such as acquiring refined grammatical categories and phrasal bracketing, segmenting words, and determining initial word-meaning associations.

The fundamental distinction between content and function words agrees with both linguistic descriptions and psycholinguistic observations. These two classes have a number of key characteristics. First, content words carry the dominant semantic load of utterances; function words are more important in marking relationships among content words in the utterance. The importance of function words for syntactic structure is encompassed in most contemporary syntactic theories.

Second, in terms of their frequency of occurrence, content words as an open class have a very large number of types whereas function words constitute a very small closed-class set. New words can easily be added to the classes of nouns, verbs, adjectives, and adverbs. Function words, in contrast, are far more stable over a long period of time. However, the frequency of occurrence of each word type exhibits the opposite pattern. The type-token ratios of function words are much higher than those for content words. That is, on average, function word types occur far more frequently than content word types. For example, the majority of the most frequently occurring 100 word types in the Kucera and Francis database [Kucera and Francis (1967)] are function words. Thus, it is observable that phrases and sentences are constructed from a small number
of function word types which occur consistently among vastly different content word types. Content words carry the dominant weight of meaning, while function words appear to serve as the structural skeleton of utterances.

Third, function words and content words are processed differently by adults. Function word errors often fail to be detected during proofreading [Rosenberg et al. (1985)]. In a recent study [Shi et al. (in press)], we showed that certain phonological processes of deletion or assimilation (such as syllable-final /t/d/ deletion when preceded by a labial in English) are more likely to occur in function words, even when frequencies are controlled. Evidence from the X-ray analysis of schwa [Yamane-Tanaka, Gick and Bird (2004)] suggests that even though schwas in function words and content words are judged to be auditorily equivalent by an observer, only the schwa in function words was found to be truly neutral in the tongue configuration, while the schwa in content words involved tongue root retraction.

Fourth, research into language acquisition has shown a different pattern for function words and content words [e.g., Brown (1973), Radford (1990)], with function words typically missing from children’s early production in many languages. As will be discussed later, the lack of function words in early production does not imply that these words are not represented and processed by infants. It does demonstrate, however, that the distinction between function words and content words exists at some level of psychological processes even in infants.

The idea that phonological characteristics may signal syntactic information has been suggested in adult language processing [Cooper and Paccia-Cooper (1980)], and was extended to the theory of prosodic bootstrapping in acquisition [Morgan (1986), see also discussions in Christophe et al. (1997)]. With respect to grammatical category distinctions per se, Kelly (1992) showed that nouns in English tend to be trochaic whereas verbs tend to be iambic, and that children are sensitive to these different sound patterns. Function words in English have been claimed to contain reduced vowels [Cutler and Carter (1987)] or described as unstressed [Gleitman and Wanner (1982)]. These arguments are primarily based on language-specific observations. Infants’ analysis of these properties likely involves a certain amount of experience with the native language so as to be able to discover the phonological cues to grammatical categories. Our model extends this idea to propose that the fundamental language-universal distinction between function words and content words is probably the first grammatical category distinction entertained by infants, and that this distinction is marked by a combination of acoustic, phonetic, phonological, and distributional cues, which exhibit a universal tendency to be reduced (although the exact manifestations of some cues must respect language-specific constraints). Furthermore, previous proposals concerning the phonological correlates of grammatical categories were largely based on the researchers’ linguistic intuitions or judgment. Our work on parental speech input therefore included systematic acoustic measures and detailed transcription analysis.

In sum, both linguistic descriptions and psycholinguistic evidence support the intuition that there is a broad, binary distinction between function and content words. Our theoretical model addresses several pertinent questions, namely, whether this distinction
is encoded in the spoken form of the input speech to infants, whether function words and content words are indeed the first two categories that the language learner derives, and if so, whether the division into two categories exerts an influence on the initial development of syntax and other aspects of language acquisition.

3. Input speech and the categorization of function words and content words

The first step in this approach was to verify whether there are acoustic and phonological cues in speech input to infants that mark the distinction between function words and content words. To test the hypothesis that function words are universally minimized in their spoken form, we recorded spontaneous maternal speech to preverbal and early verbal infants in three typologically distinct languages: English, Mandarin, and Turkish [Shi (1996), Shi et al. (1998)]. Whereas English uses limited inflections, Turkish is highly inflectional and uses rich, agglutinative functional morphemes. Mandarin, on the other hand, is an isolating language with no inflectional morphology. In our study, several types of cues were analyzed: distributional and phonological coding of the transcripts, and acoustical measures.

Based on our model, we predicted that, at the acoustic/phonetic level, all three languages should show shortened vowel duration, weaker vowel amplitude, and less pitch change for function words than for content words. Function words were expected to have a smaller number of syllables. The syllable structure was expected to be simplified such that the number of segments in the onset and coda positions is reduced or null, and the nucleus is less likely to be a diphthong. Whereas the acoustical properties of vowel duration and amplitude were expected to be language-universal in a straightforward way, the cues related to syllable number and syllable structure must respect the phonological constraints of the specific language. English, for example, permits consonant clusters in the onset and coda positions; the number of segments at these positions was expected to be reduced or null. Function words with clusters should be infrequent, although they may occur. Mandarin, on the other hand, does not permit consonant clusters in any syllable position. However, given that the syllable onset can be a consonant or null and that the coda can only be null or a nasal, function words were expected to have a tendency toward null segments in the onset and/or coda positions. The vowel F1–F2 space was also expected to be more centralized for function words than for content words. Although we only analyzed the vowel space for English (assuming that this cue might be particularly prominent because English is a stress-timing language), the reduced vowel space for function words was hypothesized to be present to some degree in all languages, as it is a property correlated with short vowel duration. Distributionally, we expected function words to be much more frequent than content words and to manifest an asymmetry with respect to whether they tended to occur in utterance-initial or -final positions.

Even though phonological properties at the syllable and word levels (e.g., number of syllables, syllable structure complexity) are subject to language-specific constraints, we
considered them to be sufficiently accessible by the infants’ perceptual system, and perhaps comparable to the language-general acoustical cues. This may not be the case for more abstract cues. That is, abstract cues may only be accessible to infants after they have learned a certain amount of the ambient language. For example, we hypothesized that the minimality of function words may be manifested in their tendency to contain unmarked segments. In our analysis, we examined whether vowels in Turkish function words were more likely to be underlyingly underspecified and thus more subject to being harmonized with the preceding vowel, and whether the tones in Mandarin function words were more likely to be unspecified and thus to exhibit the surface output of the neutral tone rule. Aside from being language-specific, such cues involve context-sensitive phonological rules and were expected to be less transparent and thus perhaps only perceptible to infants after a certain amount of experience.

The results of our analyses confirmed nearly all of the cues hypothesized for the three languages [for details, see Shi (1996), Shi et al. (1998)]. Input to preverbal infants did indeed exhibit shorter vowel duration and weaker amplitude for function words than for content words. Function words were found to consist of fewer syllables than content words; in fact, they tended to be monosyllabic. The syllable structure of function words contained fewer segmental materials than that of content words in the onset, nucleus, and coda positions, within the constraints of the native language phonology. For function words, the syllable onset and coda tended to be reduced toward nullness, and the nucleus was rarely found to contain diphthongs in languages that have a repertoire of diphthongs (such as Mandarin and English). At the abstract phonological level, the tones of Mandarin function words were significantly more likely to be unspecified (underlyingly), and the vowels in Turkish function morphemes were observed to be mostly underspecified, obtaining the remaining features from the preceding vowel. The results of distributional analyses revealed that function words occurred significantly more frequently than content words.

Our results therefore confirmed that function words were minimized in the spoken form. However, for each cue for which we found a significant mean difference between function words and content words, the distributions of the two categories showed a sizable overlap. As we described in Shi et al. (1998), this means that none of these cues alone would suffice to reliably predict the two categories. However, our model hypothesized that the correlation of the multiple cues is necessary and would be sufficient to support the sorting of the two categories. Each individual token may not, and need not, have all of the distinctive cues; a subset of cues together would indicate the membership of a token.

Given that our analyses were performed on completely spontaneous speech with no control over the linguistic contexts of the words, the results we obtained were quite robust. A study by Church (2002) attempted to replicate our findings. Church examined both spontaneous and read speech, with the targets being carefully selected. She found that unstressed vowels in function words and content words were not equivalent even when their sentence positions were identical: on average, vowels in function words were shorter than those in content words. The difference in vowel duration would be
still greater if we considered the fact that the sentence-level stress of function words is more likely to be weak than that of content words.

Having established that input speech across languages contains acoustical, phonological, and distributional cues supporting the function-word/content-word distinction, we wondered whether a language learner can derive the two categories on the basis of these cues. We first examined this question using unsupervised neural network simulations [Shi (1996), Shi et al. (1998)]. After training with a random set of words (i.e., each represented by a vector of cues that had previously been analyzed), the networks successfully classified novel words in their appropriate categories (at about 85% accuracy, a level sufficient for initial entry into the system). To test whether the cues that were generally common across languages were sufficient to derive the categories, we removed the language-specific, abstract cues (i.e., the neutral tone rule in Mandarin and vowel harmony in Turkish) from the training. The networks learned to categorize new words with comparable success. More interestingly, we found that the networks not only generalized the learned categories to words uttered by new speakers, but also to another language which they had never previously been exposed to. In other words, networks trained with words from one language were able to categorize words from another language. This suggests that the learning of the fundamental categories of function words and content words can be driven by language-universal perceptual cues.

Subsequently, we conducted experiments with newborn infants to investigate whether they were capable of categorizing function words and content words on the basis of the perceptual cues found in the input [Shi, Werker and Morgan (1999)]. In a High Amplitude Sucking Paradigm, 1- to 3-day-old infants were habituated to English words from one of the two categories and tested on novel words from either the same category or the other category. Infants showed significantly greater recovery in their sucking responses when there was a category change, indicating that they were making a categorical discrimination. This result held even when the cues of frequency of occurrence and number of syllables were balanced across the two categories. Furthermore, we compared the performance of infants who heard English prenatally versus those who heard only other languages prenatally (in nearly all cases, an Asian language). Both groups of infants responded equivalently, showing a categorical discrimination. These results suggest that the categorization must have been based on language-universal distinctive cues. Either infants’ prenatal experience with certain universal cues led to their categorical responses, or humans possess innate perceptual mechanisms that allow them to respond to universal cues even at birth.

4. Function words and language acquisition

What significance does the initial bifurcation of words into the two broad universal categories of function words and content words have for language acquisition? Our model is concerned with a number of early language-learning tasks: bootstrapping into refined

---

1 The study with newborns was limited to the use of English stimuli.
grammatical categories and rudimentary phrase structure, optimizing word segmentation, and facilitating the learning of word meaning.

At the very least, the bifurcation of words into function words and content words may considerably lower the computational load in infants’ analysis of input. It may significantly improve their learning efficacy by narrowing the scope of the analysis that they must perform. This initial division can assist with the discovery of refined grammatical categories such as nouns and verbs. Due to the high type/token ratio of function words and their regular positions in phrases and sentences, the same function word/morpheme is typically followed or preceded by many different content words of the same class, e.g., determiners are followed by different nouns. Mintz et al.’s (2002) work showed that such distributional regularities do indeed hold for nouns and verbs in input speech to English-learning children; function words play an important role in revealing the category membership of the neighboring word. Recent research with German infants [Hoehle et al. (in press)] showed that function words are involved in the early categorization of nouns.

If canonical semantic properties serve as the possible route to refined grammatical categories [Pinker (1984)], then having a rudimentary dichotomy of function words versus content words greatly constrains the possible mapping of meanings to the corresponding word forms. Indeed, we showed in a cross-linguistic study using a preference looking procedure [Shi and Werker (2001)] that infants of 6 months old selectively attended to content words [whereas the newborns described in Shi et al. (1999) did not], and that this preference was acoustically and phonologically based rather than a result of possible familiarity with specific content words [Shi and Werker (2003)]. Consistent with these results, the earliest word forms that infants are able to segment and memorize have been shown to be content words [e.g., Jusczyk and Aslin (1995), Mandel, Jusczyk and Pisoni (1995), Jusczyk and Hohne (1997), Tincoff and Jusczyk (1999)]; infants’ initial mapping of meaning to word forms as evidenced in perceptual experiments typically involves content words [e.g., Werker et al. (1998)]; and their early comprehended and produced vocabulary contains primarily content words [e.g., Brown (1973), Fenson et al. (1994)]. The fact that early learning of word meaning tends to involve content words could be due to the more transparent meaning-to-word relation of these words [Gillette et al. (1999)]. Nevertheless, the acoustically and phonologically based attentional preference for content words over function words that we observed in 6-month-olds is likely to occur first, perhaps before the onset of associating meanings to individual words. In any case, the ability to focus on one category at a time helps constrain the mapping of meanings to a smaller set of words and may facilitate word learning.

Infants’ ability to selectively attend to content words does not imply that they ignore function words. On the contrary, function words may play a crucial role in early language acquisition. In the case of word-meaning mappings, for instance, the occurrence of a determiner before a word may signal to infants that this word refers to an entity. Function words may also assist in the initial parsing of phrases. Although prosodic cues may serve as indicators of sentential and clausal boundaries [Cooper and Paccia-Copper
and infants are sensitive to such information [e.g., Hirsh-Pasek et al. (1987)], phrasal boundaries are not reliably marked by prosody. Given that any given phrase has two boundaries, the boundary that does not coincide with a clausal or sentential boundary is usually not marked prosodically. For example, the NP (noun phrase) the cat in the sentence The dog saw the cat is only prosodically marked at its end as this coincides with the end of the sentence. In this case, the function word may prove to be an important boundary marker to complement prosodic marking, potentially leading to successful bracketing of the phrase at both ends. This has been demonstrated in experiments with artificial language learning by adults [Morgan, Meier, and Newport (1987)]. Green (1979) showed that the grammar of an artificial language is unlearnable if no functional markers are present. Studies are being conducted in our laboratory to test the contribution of function words to phrasal parsing by infants.

There is evidence that infants are sensitive to function words in early language processing. Gerken and colleagues [Gerken, Landam and Remz (1990), Gerken and McIntosh (1993)] showed that 2-year-olds comprehended and produced utterances containing real, grammatical functors better than those either with functors replaced by nonsense functors or with ungrammatical functors. Using time-locked eye movement measures, Zangl and Fernald (2003) showed that English-learning infants as young as 18 months of age were sensitive to function morphemes in online sentence processing. Two-year-old Dutch-learning infants have been shown to have some knowledge of grammatical gender [Johnson (2004)]. Recent work from a few laboratories suggests that even preverbal infants are sensitive to function words [Shady (1996), Schafer et al. (1998), Hoehle and Weissenborn (2003), Shi, Werker and Cutler (2003, 2004a)].

In our recent work, we tested the hypothesis that function words, especially very frequent function words, may facilitate the segmentation of content words in continuous speech in early infancy. Word segmentation is necessary since most speech directed to infants is not in citation form. Approximately 90% of utterances in our recordings of motherese in Mandarin, English, and Turkish were multiword utterances. In a study of English infants [Shi et al. (2004b)], 11-month-olds segmented a monosyllabic content word preceded by the frequent functor the better than when a content word was preceded by a nonsense functor. After becoming familiar with the sequences “the + content word 1” versus “nonsense functor + content word 2,” infants listened significantly longer to isolated presentations of “content word 1” than to “content word 2” during the test phase. That is, they segmented the syllable better when it followed the than when it was preceded by a segmentally modified nonsense functor (it is therefore possible that they may have treated the sequence “nonsense functor + content word 2” as one single disyllabic word, with “content word 2” as the second syllable of the word). No such difference in segmentation was found for sequences including an infrequent functor versus a nonsense functor. This shows that only the function word the facilitated the segmentation of an adjacent content word.

The effects of function words on early language acquisition and processing discussed above imply the assumption of a prior segmentation of the word forms of function words
from continuous speech. This is an empirical question, but a highly plausible one. Given infants’ remarkable ability to use statistical-distributional properties to segment the speech stream into word-like units [Saffran, Newport and Aslin (1996)], frequently occurring function words in highly varying contexts are likely to be segmented at quite an early age. Hoehle and Weissenborn (2003) showed that German-learning 7- to 9-month-old infants segmented function words in continuous speech. Behavioral experiments [Shady (1996)] and ERP measures [Schafer et al. (1998)] using synthetic speech revealed that English-learning 10.5- and 11-month-old infants can detect function words in utterances. In our experiments using a preference procedure [Shi et al. (2003, 2004a, under review)], we presented 8-, 11-, and 13-month-old infants with phrases containing real function words versus nonsense function words. We found that 11-month-old infants showed an emerging recognition of function words as a class, and 13-month-olds showed a robust recognition. Furthermore, because the nonsense function words in our stimuli were minimal modifications of the real functor counterparts (with prosody unchanged), our results also suggest that infants’ representation of the functors they recognized was phonetically well specified.

In these studies on infants’ perception of function words [Shi et al. (2003, 2004a, under review)], we considered the idea that the word forms of certain function words are stored in infants’ memory, i.e., an emerging mental lexicon (so that they may exert an influence on the linguistic analyses discussed above), and that the function words with extremely high frequencies of occurrence are likely the earliest candidates in the initial lexicon. In our view, these items are unlikely to have any meaning in the early lexicon. Instead, they are stored only as pieces with phonetic representations. Our finding that segmentation of an adjacent content word was only facilitated in the context of a very frequent functor, but not in the context of an infrequent or nonsense functor [Shi et al. (2004b)], suggests that it was the very frequent functors previously stored in infants’ emerging mental lexicon (or some prototypical phonological representation, e.g., a schwa, for certain function words) that exerted the facilitation effect. Note that every real versus nonsense functor was presented equally frequently during the experiments, indicating that the frequency effect observed must have been due to the unequal frequencies of occurrence in the natural input that the infant had heard.

Developing an early lexicon which includes at least some spoken forms of function words/morphemes, regardless of whether their phonetic representations are well specified, may impact on various important aspects of early language acquisition. The stored function words could facilitate infants’ discovery of phrasal structures, the derivation of

---

2 In the Saffran et al. study, infants were presented with a monotonic synthetic speech stream consisting of nonword syllables. Only information on transitional probabilities between syllable sequences was kept and manipulated. A sequence of syllables with a high transitional probability was more likely to be a word-like unit than a sequence containing syllables with a low transitional probability. To illustrate with a real language, for the phrase *joli visage*, the transitional probability between *jo* and *li* would be higher than that between *li* and *vi*; this is because *li* can be followed by many different words such as *joli chat*, *joli jardin*, etc., while *jo* and *li* co-occur more consistently. Saffran et al. found that infants were able to rely on this statistical information to segment the “words” in the artificial language.
a refined set of grammatical categories, word segmentation, and the learning of word meanings, as discussed in this section.

5. Conclusions

In this chapter, I addressed the question of how infants initially break into language. In particular, I argued that infants may first establish the syntactic categories of function and content words, and that these initial categories may then assist them in their subsequent analyses of the language input.

In our model, we hypothesized that input speech to infants contains universal acoustic, phonetic, and phonological cues to the most fundamental, language-universal, superordinate distinction between function words and content words. More specifically, function words were hypothesized to be universally minimized in the spoken forms, although the manifestations of this minimality must respect language-specific constraints at the more abstract phonological level to some degree. We further proposed that infants are capable of categorizing function words and content words on the basis of these perceptual cues. As discussed in Section 3, data from cross-linguistic studies of infant-directed speech, neural network simulation of the categorization of the two categories, and newborn infants’ categorical discrimination of the two categories all support the above hypotheses.

Another component of our model includes the hypothesis that the initial bifurcation of words into function words and content words may facilitate or optimize various important language-learning tasks, including the derivation of refined grammatical categories and phrasal structures, learning of word meanings, and word segmentation. This chapter presents arguments in favor of each of these aspects as well as the existing empirical results that support some of them.

Finally, the model was extended to consider the idea of an impoverished initial mental lexicon at the preverbal stage that includes highly frequent function words; it is these functors that first play a facilitation role in the language-learning tasks discussed above. The results of our recent experiments with 8- to 13-month-old infants support this proposal. It is argued in Section 4 that frequent function words in the emerging lexicon contain only phonetic forms. These stored function words, which contain no meaning, primarily play the role of assisting infants’ early language analyses such as refined grammatical categories, rudimentary phrasal bracketing, word segmentation, and word-meaning mapping.

Acknowledgment

The research reported in this chapter was supported in part by the National Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds québécois de la recherche sur la société et la culture.
References


Radford, A. (1990), Syntactic Theory and the Acquisition of English Syntax (Blackwell, Oxford).


This page intentionally left blank
ACQUIRING AUDITORY AND PHONETIC CATEGORIES*

MARTIJN GOUDBEEK, ROEL SMITS, AND ANNE CUTLER

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

DANIEL SWINGLEY

Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, University of Pennsylvania, PA, USA

Contents

Abstract 497
1. Introduction 498
2. Testing category learning 500
3. Learning of nonspeech categories 502
4. Learning of speech categories 506
5. Conclusion 510
References 511

Abstract

Infants’ first steps in language acquisition involve learning the relevant contrasts of the language-specific phonemic repertoire. This learning is viewed as the formation of categories in a multidimensional psychophysical space. Categorization research in the visual modality has shown that adults are unable to learn multidimensional categorization problems without supervision. The success of infants in acquiring a phonetic system suggests that formation of multidimensional categories should be more tractable in the auditory modality. We describe experiments investigating adult learning of multidimensional speech and nonspeech categories. These experiments revealed that the degree of difficulty is actually significantly greater than that observed in the visual modality. Despite comparable methods, our results differ from those in visual category learning: feedback that is effective for visual categories is not effective in the auditory modality. Attending to more than one dimension in auditory category formation is possible for adult listeners, but very hard.

* Part of this research was carried out with financial support from the Dutch Scientific Research Council. We further thank Keith Kluender, University of Wisconsin, Madison, for financial and other assistance with the speech experiments.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved

497
1. Introduction

Both infants and learners of a second language are faced with the problem of correctly categorizing the sounds of a new language. This task breaks down into two problems for the listener. The first is how to acquire the categories in the first place. Starting without any knowledge of the sounds of a language, how do listeners extract the categories from the language around them? The second problem is the mapping of incoming acoustic material to the categories, once they have been acquired. In this chapter, we are primarily concerned with the first question. The online mapping and representation of auditory categories is investigated by, for example, Smits, Sereno, and Jongman (in press). Learning the sounds of a second language presents the listener with the additional problem of a psychophysical space that is already structured by the first language [Best, Roberts, and Sithole (1988)]. The basic process of acquiring the categories of a second language is the same as with the first, although the first language will always interfere with listening to a second language [Cutler and Broersma (2005)].

Our approach to the issue of the acquisition of an auditory category is similar to that applied in visual category learning studies [Nosofsky (1992), Ashby and Maddox (1993)]. Analogous to the studies investigating visual category learning, our work concerns perceptual categories, i.e., categorization in a psychophysical space with continuous dimensions. This may be different from categorization learning for “higher-order” categories involving dichotomous features such as “shape of the head” (round versus oval) or “long versus short legs” [Minda and Smith (2001, 2002)]. We assume that when listeners hear a sound, this sound is evaluated in a number of dimensions and mapped onto a point in a multidimensional psychophysical space. Repeated exposure to sounds originating from distinct categories will thus lead to “clouds” of points in the space. If, after a period of exposure, several more or less distinct clouds emerge, the subjects may start to associate the clouds with different categories. Such a conceptualization was implemented by Behnke (1998) in a neural network that recognizes patterns in a phonetic map; Kornai (1998) likewise constructed a neural network modeling the data of Peterson and Barney (1952).

The capacity of newborns to discriminate speech sounds is remarkable. In the first few months of life, they discriminate a wide range of speech-sound contrasts, but over the course of the first year they start to conflate similar sounds if those sounds do not make a phonological contrast in their native language [see, e.g., Aslin, Jusczyk, and Pisoni (1998) or Jusczyk (1997) for reviews]. Several studies have found decrements in non-native consonant discrimination by the age of 12 months [e.g., Werker and Tees (1984)] and analogous decrements in non-native vowel perception even earlier [Kuhl et al. (1992), Polka and Werker (1994)].

The above description holds for the discrimination of speech sounds, but not necessarily for their categorization. Discrimination involves the comparison of at least two recently heard stimuli, but categorization involves the comparison of a stimulus to a stored category representation. Category learning entails that certain stimuli that may be perceptually distinguishable are associated with a single category, and thus become,
in some sense, identical. Eventually, such category learning may cause the once easily
discriminable stimuli to become less discriminable or even indiscriminable [Goldstone
(1994)]. Thus, the learning of categories may at least partly explain the loss of infants’
abilities to discriminate between phonemes that are not contrastive in their native lan-
guage. Infants’ phonetic categorization abilities are not simply equivalent to their dis-
criminatory abilities, because true acquisition of a phonetic category requires not
simply the failure to discriminate between members of that category, but also the abil-
ity to assign the appropriate category label to these members. Perception of similarity,
as evidenced, for example, by the studies of Kuhl [Hillenbrand (1983), Kuhl (1985),
Cameron Marean, Werner, and Kuhl (1992)], is undoubtedly an important aspect of cat-
egorization. However, the literature does not yet, in our opinion, motivate a complete
account of the relationship between stimulus discrimination and stimulus categorization
by infant listeners.

The acquisition of speech categories by infants happens in a wholly unsupervised
way. Simply by being exposed to speech, and without formal instruction, infants
acquire their native phonetic categories and lose the ability to discriminate non-native
contrasts. This process occurs too early in development to be driven by semantic con-
trasts in phonetically similar words and is therefore generally considered to result from
infants’ distributional analysis of speech. Supporting this notion, experimental studies
have shown that infants are extremely sensitive to the statistical properties of incoming
particular relevance to phonetic category learning is a study by Maye, Werker, and
Gerken (2002), in which two groups of infants were exposed to /da/–/ta/ stimuli in a
preferential looking procedure. One group listened to a unimodal distribution, the other
to a bimodal distribution of stimuli, encouraging them to group the sounds into two cat-
egories. Infants who listened to a bimodal distribution in the training phases listened
longer to Alternating (two different stimuli) trials compared to Non-Alternating (the
same stimulus repeated) trials in the test phase. Infants exposed to a unimodal distribu-
tion did not show this differential looking [Maye et al. (2002)]. Similar sensitivity to
distributional properties has been found for adults [Maye and Gerken (2000, 2001)].
Thus, Maye and colleagues have shown that both infants and adults are capable of
learning unidimensional categories without supervision. As Pierrehumbert (2003)
points out, however, it is not clear whether this result would generalize to less salient
phonetic dimensions or to noisier training conditions.

We assume that statistical learning lies at the heart of phonetic category acquisition;
aquiring these categories is simply equivalent to recognizing the statistical patterns
present in speech. Statistically based category learning has previously been most inten-
sively studied with respect to visually presented stimuli, and our experimental approach
uses methods and insights from such visual categorization studies.

The general methodology is described in Section 2. In Section 3, we describe a study
investigating the learning of two nonspeech categorization problems, one with a unidi-
mensional category distinction and one with a multidimensional category distinction.
In the third section, the learning of speech categories is investigated, again with a
unidimensional and a multidimensional category distinction. The last section summarizes and discusses the presented experiments.

2. Testing category learning

In a two-dimensional psychophysical space we defined two categories as two-dimensional probability density functions, each represented by a “cloud” of points. We varied the relevance of each of the dimensions by manipulating the orientation of the clouds. For example, exposure to the structure in the top left cell in Figure 1 should encourage subjects to categorize using only the vertical dimension (dimension 1), whereas exposure to the structure in the bottom left figure should encourage subjects to use only dimension 2. Exposure to the structures in the right-hand column should increase the tendency to use both dimensions in categorizing, because the use of only one dimension would lead to many incorrect categorizations.

The experiments described in this chapter used a single basic procedure, in which a training phase was followed by a test phase. In the training phase, subjects heard

![Fig. 1. Four category structures in a two-dimensional space.](image)
stimuli drawn from two probability density functions. Solving the categorization problem required the use either of one dimension or of both dimensions simultaneously, depending on the experiment. If listeners chose a unidimensional criterion, they would assign all stimuli below a criterial value on that dimension to one category and all stimuli above that value to the other category. To categorize via more complex rules, such as the diagonal line that separates the categories in the right-hand column of Figure 1, they would have to use more dimensions. After the training phase, the listeners entered a test phase, which was intended to assess the subjects’ categorization behavior across the psychophysical space. Subjects heard and categorized stimuli that were drawn from an uniformly spaced grid, in which distributional information was no longer present (see the right panel of Figure 2).

The experiments used supervision, i.e., feedback, in the training phase. This contrasts with the infant situation where supervision is absent, but allows for a direct comparison with the visual perception studies and models of second language learning in adults; supervised learning of unidimensional and multidimensional auditory categories has seldom been studied and may thus inform models of the learning of second language categories, a process which, at least in part, is often supervised.

One hazard in applying the above methodology to speech perception experiments is that subjects may use the representation of the sounds of their native language in solving the categorization problem [Best and Strange (1992)]. We therefore first presented listeners with nonspeech stimuli that have been shown to be dissimilar to any known sound categories [Smits et al. (in press)]. Combining experiments with nonspeech stimuli as an analog for speech with experiments with speech stimuli also addresses the nature of the exact relationship between speech and nonspeech stimuli. Speech and nonspeech could be analyzed by partly different perceptual systems, but they could also to a large extent be perceived by the same system. In the set of experiments with speech stimuli, we minimized the influence of existing native categories by using vowel stimuli that would map to relatively “empty” areas of the subjects’ native vowel space.

![Fig. 2. Category structures of conditions 1 and 2 (left and middle panels) and the test phase (right panel) for both the speech and nonspeech experiment.](image-url)
3. Learning of nonspeech categories

In our first study, subjects had to learn either a unidimensional (condition 1) or a multidimensional category structure (condition 2). In condition 1, only duration was important for creating the category distinction (see the left-hand panel of Figure 2), and subjects had to learn to ignore the other dimension, whereas in condition 2, both dimensions were relevant (see the middle panel of Figure 2). The category learning process was supervised: subjects received trial-by-trial feedback. In the test phase, listeners categorized stimuli from the equidistantly spaced grid (see the right-hand panel of Figure 2) without receiving feedback.

The stimuli were inharmonic tone complexes filtered by a single resonance, 112 in each category. We used inharmonic sounds because harmonic complexes tend to be easily associated with speech. The two dimensions of variation in the experiment were the frequency of the spectral peak at which the sound complex was filtered (“formant frequency”) and the duration of the stimulus (“duration”). Many previous investigations have shown these dimensions to be very important in the perception of vowel sounds [e.g., Ainsworth (1972), Peterson and Barney (1952)]. Detailed descriptions of the stimulus construction can be found in Smits et al. (in press).

Table 1 lists the properties of the stimuli. The distributions were defined in a “perceptual scale” spanned by psychological scales. The psychophysical space commonly accepted for the perception of frequency is the equivalent rectangular bandwidth (ERB) scale [Glasberg and Moore (1990)]. The ERB is derived from frequency according to the following transformation, where $f$ refers to frequency in Hertz:

$$ERB = 21.4 \times 10 \log(0.00437 * f + 1)$$

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Step size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>50.1 DUR</td>
<td>47.6 DUR</td>
<td>52.6 DUR</td>
<td>0.5 DUR/step</td>
</tr>
<tr>
<td>Frequency</td>
<td>18.8 ERB</td>
<td>17.6 ERB</td>
<td>20.0 ERB</td>
<td>0.2 ERB/step</td>
</tr>
</tbody>
</table>
Psychological duration $D$, the psychophysical counterpart of duration, was calculated from stimulus duration $t$ by Smits et al. based on data provided by Abel (1972) as follows:

$$D = 10 \log t$$

To ensure that both dimensions would be equally salient and discriminable, they were normalized using their respective just noticeable differences (Weber fractions). For formant frequency, this is equivalent to 0.12 ERB in the relevant frequency region [Kewley-Port and Watson (1994)], while for duration pilot experiments of Smits et al. (in press) and our subsequent piloting with stimuli varying in duration and frequency indicated that 0.25 $D$ resulted in a discriminability comparable to that of 0.12 ERB. These values were used to establish the range of variation of the stimuli, so that the difference between the lowest and the highest value on both the frequency and the duration dimension equaled 20 just noticeable differences. Twelve subjects, students from the University of Nijmegen without a history of hearing problems, participated in return for a small payment. They were seated in a soundproof booth in front of a computer screen and a response button box. In the training phase, they listened to 448 stimuli, two times 112 from each of the two categories they were attempting to learn. In condition 1, the stimuli varied irrelevantly in one dimension, duration, and relevantly in the other dimension (frequency of the spectral peak). In condition 2, the stimuli manifested relevant variation in both dimensions (see the left and middle panels of Figure 2). The subjects’ task was to assign each stimulus to group A or B, using the two-key button box. When their categorization was correct, the monitor displayed the Dutch equivalent of “RIGHT” in green letters, for 700 ms; when the categorization was incorrect, the monitor displayed the Dutch equivalent of “WRONG” in red letters for 700 ms. After the visual feedback disappeared, the next stimulus was presented. Subjects were offered a brief pause halfway through the training phase, and again between the second part of the training and the test phase. In the test phase, subjects categorized sounds from the test continuum (right panel of Figure 2) as belonging to group A or B, this time without receiving feedback regarding their response.

Subjects’ responses were analyzed using a logistic regression technique. Figure 3 shows the mean $\beta$-weights from the logistic regression separately for the first half of the training (“Training Part 1”), the second half of the training, (“Training Part 2”) and the test stimuli (“Test”). The $\beta$-weight indicates the extent to which a factor (dimension) can be used to fit the data with a logistic function. We divided the results for the training phase into two parts to probe for learning during the training phase. Table 2 displays the full results of the logistic regression analysis. Subjects clearly learned to use the relevant dimension in the training phase, and they retained this ability in the test phase.

In the individual analysis of each subject, the size of the $\beta$-weight may or may not be significantly different from zero. When it is not significantly different from zero, subjects did not make any significant use of the dimension in question. The columns labeled “Uni” and “Multi” show this information. A subject who uses both dimensions is counted under “Multi,” while a subject who uses only one dimension to a significant extent is counted under “Uni.” Subjects who use no dimension are not counted.
Condition 1 presents a clear picture of learning of a category structure with one relevant dimension of variation. Subjects discover the relevant dimension early in the training phase, but their performance improves throughout their training, as evidenced by the increasing $\beta$-weights. The difference in the $\beta$-weight for duration between the first and
second part of the training is significant \( F(2,33) = 6.895, p < 0.05 \). Post hoc analyses of the \( \beta \)-weights with the Student–Newman–Keuls test show that Training Part 2 and the Test form a homogeneous subset \( p < 0.05 \), while Training Part 1 differs from both Training Part 2 and Test. The difference in the \( \beta \)-weight for frequency between the first and second parts of the training is not significant \( F(2,33) = 1.725, \ p > 0.2 \).

The results of condition 2 are harder to interpret. Figure 4 displays the \( \beta \)-weights for both dimensions, while Table 2 shows the values of the \( \beta \)-weights and the number of subjects preferring a unidimensional versus a multidimensional solution. As the size of the \( \beta \)-weights indicates, performance is not very good compared with condition 1. There are significant effects of training in condition 2, for frequency \( F(2,33) = 13.223, \ p < 0.05 \) and duration \( F(2,33) = 3.901, \ p < 0.05 \). The individual data show that in Training Part 2, 6 out of 12 subjects use both dimensions significantly in their categorization. This learning does not, however, generalize to the test phase, when only one subject still uses a multidimensional strategy.

Post hoc analyses of the \( \beta \)-weights for the dimension duration show that Training Part 1 and Training Part 2 are one homogeneous subset, as is the test phase. So duration is not used more in the second part of the training phase, compared to the first. In the test phase, the \( \beta \)-weight for duration is different from that in the training. Without the information provided by the supervision and the category structure, subjects start using duration in their categorization. For frequency, the situation is different; the post hoc analyses reveal that subjects learn to use frequency during the training phase; Training Part 1 and Training Part 2 are different (heterogeneous) subsets according to the Student–Newman–Keuls test. Training Part 1 and Test do not differ significantly, which indicates that the learning does not generalize to the test phase.

![Fig. 4. \( \beta \)-weights for formant frequency and duration in the three phases of the nonspeech experiment, condition 2: relevant variation in both dimensions (formant frequency and duration).](image-url)
Comparing condition 1 and condition 2 shows that real learning took place in condition 1, with performance improving during the training and reaching ceiling in the test. In condition 2, on the other hand, only about half of the subjects learned to use both dimensions in the training. In the test phase of condition 2, however, subjects almost invariably changed their categorization strategy, choosing one dimension to categorize the stimuli.

This study showed that, as expected on the basis of the evidence from visual category learning, participants who received feedback about their categorizations readily learned distributionally defined categories, and they were able to ignore irrelevant variation. However, even with feedback, performance in learning to use two distinct dimensions for separating categories defined over both dimensions was difficult. In the next section, the generality of this result was evaluated by testing listeners on synthesized tokens of nonnative vowel sounds.

4. Learning of speech categories

Our investigation of speech category learning was directly analogous to the nonspeech study. To assess possible pre-existing categorization tendencies for these speech stimuli, the experiment began with a pretest phase. Because the pretest stimuli and procedure were identical to the test phase, any difference between the pretest and the test phase can be attributed to the effect of the training phase.

The stimuli were synthesized versions of the Dutch front vowels /ø/, /y/, and /Y/ as in the Dutch words feut (/føt/, ‘freshman’), fuut (/fyt/, ‘grebe’), and fut /fYt/, ‘energy’). The differences between these vowels can be described by the duration and frequency of the first formants [Nierop, Pols, and Plomp (1973), Pols, Tromp, and Plomp (1973)]. The first formant of both /ø/ and /Y/ is approximately 441 Hz (about 8 ERB), while that of /y/ is approximately 305 Hz (about 10 ERB). As for duration, /Y/ and /y/ are short vowels of approximately 98 ms (about 46 D) while /ø/ is longer, at an average of 161 ms (about 51 D). All stimuli were generated using the PRAAT speech synthesis program [Boersma (2001)]. With the exception of frequency, all formants were kept constant for all stimuli (see Table 3). The values for the training stimuli were obtained by random sampling from two distributions. Careful listening by Dutch listeners (the first and second authors) revealed that the means of the categories still qualified as acceptable examples of the two Dutch vowels. In condition 1, the categories were defined by separation in the dimension of duration, with equivalent irrelevant variation in first formant frequency. In condition 2, the categories were defined by separation in both dimensions – duration and formant frequency.

Summary statistics for the stimuli are presented in Table 3. The test stimuli had the same design as in the nonspeech study: a grid with dimensions of equal psychological range.

All subjects were students from the University of Wisconsin, Madison. Some participated in exchange for course credit, while others were paid for their participation.
All were native speakers of English and did not speak another Germanic language or any other language with front-rounded vowels. Before the experiment, subjects filled out a consent form. In condition 1 there were 10 subjects, and in condition 2 there were 18 subjects.

Subjects were seated in a soundproof booth and were provided with a response box with four buttons and a light above each button. The experiment consisted of three phases; a pretest phase, a training phase, and a test phase. In the pretest and in the test phase, subjects listened to 196 (4 × 49) stimuli from the speech test continuum and categorized them as belonging to group A or B, without receiving feedback. In the training phase, subjects listened to 448 (2 × 2 × 112) speech training stimuli. The stimuli were presented in randomized blocks of 224. Training Parts 1 and 2 were analyzed separately. As in the nonspeech experiment, the difference between the conditions was in the number of relevant dimensions. In condition 1, the stimuli from a structure varied irrelevantly in one dimension (formant frequency) and relevantly in another (duration). In condition 2, the stimuli manifested relevant variation in both dimensions (see the left and middle panels of Figure 2). Subjects had to assign a stimulus to the leftmost or the rightmost of four buttons. A light above the button indicated which button was the correct response. If the light turned on, the categorization was correct; if not, the response was incorrect. The light stayed on for 400 ms, after which the next trial began.

Again, the results for condition 1 presented a clear picture (see Figure 5). There was a clear learning effect for duration ($F(3, 36) = 12.365, p < 0.05$). Subjects learned to use duration during the training phase, and retained this ability. There was also a small, but significant, effect for F1 ($F(3, 36) = 6.6, p < 0.05$) showing that subjects did not ignore this dimension entirely. Post hoc analyses with the Student–Newman–Keuls test

<table>
<thead>
<tr>
<th>Training stimuli</th>
<th>Category A “/ø/”</th>
<th>Category B “/ɪ/”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Means</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>Condition 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50.7 DUR</td>
<td>0.5 DUR</td>
</tr>
<tr>
<td></td>
<td>10.0 ERB</td>
<td>0.2 ERB</td>
</tr>
<tr>
<td></td>
<td>10.0 ERB</td>
<td>0.2 ERB</td>
</tr>
<tr>
<td>Condition 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>49.7 DUR</td>
<td>2.6 DUR</td>
</tr>
<tr>
<td></td>
<td>10.1 ERB</td>
<td>0.2 ERB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fixed formants</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.6 ERB</td>
<td>22.3 ERB</td>
<td>26.2 ERB</td>
<td>28.2 ERB</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test stimuli</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Step size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>48.5 DUR</td>
<td>45.85 DUR</td>
<td>50.8 DUR</td>
<td>0.5 DUR/step</td>
</tr>
<tr>
<td>F1</td>
<td>10 ERB</td>
<td>9.9 ERB</td>
<td>10.5 ERB</td>
<td>0.3 ERB/step</td>
</tr>
</tbody>
</table>
revealed that for duration there were three separable homogeneous subsets, the pretest, the training, and the post-test \((p < 0.05)\). For frequency, Training Part 1 was different from the other three parts \((p < 0.05)\). Table 4 shows that in the pretest there was no preference for using either duration or frequency in categorizing the stimuli. After the first training phase, 9 of the 10 subjects consistently used duration and this generalized to the test phase.

As in experiment 1, the results of condition 2 are harder to interpret. According to Figure 6, the average use of both conditions increases. However, because the increase is small, this effect is not significant for duration \((F(3, 68) = 0.189, p < 0.904)\) or for frequency (though the latter effect is stronger: \(F(3, 68) = 2.189, p < 0.097)\). In the test phase the learning effect washes away, and subjects hardly use duration anymore. The post hoc tests support this interpretation, as all phases of the experiment are in the same subset.

The individual subjects’ data in Table 4 show the same pattern: although subjects exhibited a preference for frequency in the pretest, this changed in the first phase of the training, where both duration and frequency were used by five subjects. Six subjects already used both dimensions simultaneously, and this increased to nine subjects in the second phase of the training. This learning effect did not generalize to the test phase, where subjects preferred to use frequency in categorizing the stimuli.

Comparing conditions 1 and 2 yields the same result as in experiment 1. In condition 1, subjects learned to use the relevant dimension, whereas in condition 2 using both dimensions was considerably more difficult. Still, half of the subjects learned to use two dimensions in the second phase of the training, and one-third of the subjects used two dimensions in the test phase.

Fig. 5. \(\beta\)-weights for formant frequency and duration in the four phases of the speech experiment, condition 1: relevant variation in one dimension (duration).
Table 4
Logistic regression results for the speech experiment for each condition

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 relevant dimension, $N = 10$)</td>
<td>(2 relevant dimensions, $N = 18$)</td>
</tr>
</tbody>
</table>

**Pretest**

<table>
<thead>
<tr>
<th></th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>0.30</td>
<td>0.14</td>
<td>3</td>
<td>0</td>
<td>0.33</td>
<td>0.10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.13</td>
<td>0.04</td>
<td>4</td>
<td>0</td>
<td>0.86</td>
<td>0.19</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Training part 1**

<table>
<thead>
<tr>
<th></th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>1.62</td>
<td>0.24</td>
<td>10</td>
<td>0</td>
<td>0.44</td>
<td>0.18</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.11</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0.96</td>
<td>0.08</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

**Training Part 2**

<table>
<thead>
<tr>
<th></th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>2.11</td>
<td>0.36</td>
<td>9</td>
<td>0</td>
<td>0.51</td>
<td>0.12</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.23</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>1.06</td>
<td>0.21</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

**Test**

<table>
<thead>
<tr>
<th></th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
<th>Mean $\beta$</th>
<th>$\sigma$</th>
<th>Uni</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>1.16</td>
<td>0.17</td>
<td>9</td>
<td>0</td>
<td>0.20</td>
<td>0.06</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Frequency</td>
<td>0.08</td>
<td>0.02</td>
<td>1</td>
<td>0</td>
<td>0.99</td>
<td>0.19</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Notes: $\beta$-weights are shown for both dimensions and the number of subjects using one (Uni) or both (Multi) dimensions significantly.

Fig. 6. $\beta$-weights for formant frequency and duration in the four phases of the speech experiment, condition 2: relevant variation in both dimensions (formant frequency and duration).
5. Conclusion

Auditory category learning for both speech and nonspeech stimuli is strikingly difficult. Our data show that listeners can relatively quickly learn a unidimensional categorization, but a categorization involving the simultaneous use of two dimensions is very hard to acquire. The unidimensional categorization was robustly acquired even though listeners had to ignore irrelevant variation in another dimension. The multidimensional categorization remained hard even though listeners had at their disposal two sources of information about the category structure, namely both the distributional characteristics of the category exemplars and feedback regarding their category judgments.

Learning of a multidimensional category structure is, we argue, the task facing infants who are learning to map the phonemes of their native language onto their psychophysical space. Yet we have shown that this task is extremely difficult for adult listeners. There are several possible explanations for this discrepancy.

First, the differences between the learning capacities of infants and adults may be greater than we had hypothesized. Adults may simply have lost the pattern recognition skills available to infants. Also, adults have already acquired a phonemic inventory, which may interfere with learning new phonetic categories or analogous auditory categories [Best and Strange (1992)]. Second, it is conceivable that explicit feedback leads subjects to attempt to formulate a verbal description or rule for use in their categorizations. Such verbal rules generally cannot separate multidimensionally varying categories, as argued by Ashby and colleagues [Ashby et al. (1998), Maddox, Ashby, and Waldron (2002), Maddox, Bohil, and Ing (2004)]. Third, infants receive a lot more exposure in acquiring their phonetic categories. Perhaps auditory category learning, unlike visual category learning, requires a very large number of exposures.

All current proposals for how infants spontaneously learn phonetic categories are distributional learning accounts in which infants are argued to perform statistical clustering over large numbers of isolated tokens of speech sounds. Experimental evidence with infants is compatible with this notion in broad outline, but in fact surprisingly little is known about the learning of auditory categories, either in infancy or in adulthood. The present experiments used techniques borrowed from related studies in the visual modality, presenting subjects with extensive exposure to distributionally defined categories with dimensions of variation known to be discriminable. Even with stimuli varying along two dimensions over a range of 20 just noticeable differences, though, and with supervised training, category learning performance in adults for both speech and nonspeech stimuli was poor and inconsistent, and collapsed soon after the close of training.

From a methodological point of view, of course, it is encouraging that the results for the nonspeech stimuli and the speech stimuli were so similar. This supports our assumption that the nonspeech stimuli we used were good counterparts of the speech stimuli. The former were not recognized as being speech, yet they were analyzed in a similar manner. However, the subjects’ overall poor performance in auditory category acquisition remains puzzling. What learning occurred in our study seemed, moreover, rather fragile. In the test phase of our experiments, we observed in all conditions, and for both nonspeech and
speech, that subjects “unlearned” the categorization rule they had previously mastered. We suggested above that this resulted from the stimulus configuration with which listeners were presented in the test phase. The use of a grid with uniformly spaced stimuli to assess the psychophysical space of a listener is, in fact, a well-known technique in the field of phonetics and phonology. The lack of information in the distribution of the stimuli is intended to prevent subjects from changing their categorization tendencies. However, this is not what happened; our listeners apparently picked up on the fact that in the test phase both dimensions were equally relevant, and altered their categorizations to suit.

Studies of auditory perceptual learning with respect to already known phonetic categories have shown that adult listeners exhibit remarkable flexibility in adjusting the boundaries of the phoneme categories of their native language [Norris, McQueen, and Cutler (2003), Evans and Iverson (2004), Eisner and McQueen (2005)]. Such adjustments enable listeners to adapt quickly to new speakers and new dialects; eventually, perceptual adjustments of this nature also underlie pronunciation changes across whole language communities. The listeners in our experiments seemed to maintain analogous flexibility towards the use of auditory information in the input; where they performed poorly was in converting the information into confident categorical decisions. The categories which were most speech-like, in that they were defined by multidimensional variation, were the hardest for these adult listeners to acquire. By contrast, the unidimensionally defined categories were reasonably well learned despite substantial irrelevant variation in a second dimension. These results thus point to significant gaps in our current understanding of auditory category learning, and of how infants acquire phonetic categories. The next step is, clearly, to endeavor to close these gaps.

References


Ch. 22: Acquiring Auditory and Phonetic Categories


Chapter 23

SYNTACTIC CATEGORIES IN SECOND LANGUAGE ACQUISITION

LYDIA WHITE

McGill University

Contents

Abstract 515
1. Introduction 516
2. Lexical and functional categories 516
3. Lexical categories in L2 acquisition 517
4. Functional categories in acquisition: Issues of evidence 519
5. Functional categories in the L2 initial state and in L2 development 522
   5.1. Morphology-before-syntax 522
   5.2. Syntax-before-morphology 523
6. Acquiring versus losing categories and features 524
7. Discussion 529
References 530

Abstract

This chapter discusses how syntactic categories (lexical and functional) and properties associated with them are acquired and represented in the interlanguage grammars of second language (L2) speakers. The following issues are discussed: (i) the nature of the evidence required to determine whether categories are represented; (ii) whether or not interlanguage grammars initially lack certain categories altogether; (iii) the influence of the first language (L1), namely whether or not L2 learners simply adopt for the L2 those syntactic categories that are represented in the L1; (iv) acquisition of new categories (required to represent the L2 but not the L1) and loss of categories (required to represent the L1 but not the L2).
1. Introduction

In the field of second language (L2) acquisition, it has long been assumed that the underlying linguistic competence of learners or speakers of a second language (L2) is represented in terms of a complex linguistic system, commonly referred to as an interlanguage grammar [following Selinker (1972)]. Current generative linguistic perspectives on L2 acquisition make explicit claims about the nature of interlanguage grammars. This chapter discusses how syntactic categories (and properties associated with them) are acquired and represented in L2 acquisition. Debate centers on the following issues: (i) the influence of the first language (L1), namely whether or not L2 learners simply adopt for the L2 those syntactic categories that are represented in the L1; (ii) whether or not L2 learners can draw on the full range of categories made available by Universal Grammar (UG), when the L1 and L2 differ as to the categories that they instantiate. We will also consider whether or not interlanguage grammars initially lack certain categories altogether and whether it is in some sense harder to “lose” categories than to acquire them.

2. Lexical and functional categories

In linguistic theory, syntactic categories are divided into two types: lexical categories, which are categories with semantic content, such as verb (V), noun (N), adjective (Adj), adverb (Adv), and preposition (P); and functional categories, which represent grammatical properties. Typical functional categories are complementizer (Comp or C), inflection (Inf1 or I), tense (T), agreement (Agr), negation (Neg), determiner (Det), and number (Num). Functional categories are made up of bundles of formal grammatical features (such as tense, number, person, gender, case, definiteness, and specificity); indeed, it is currently assumed that features rather than categories are primitives in the theory.

Given that many different functional categories have been proposed in the recent literature [e.g., Pollock (1997), Rizzi (1997), Cinque (1999)], it seems unparsimonious to assume that all categories are represented in all languages. Many researchers consider that grammars vary crosslinguistically in terms of the categories and/or features that they instantiate. This has been argued for verbal projections [e.g., Iatridou (1990), Webelhuth (1995), Thráinsson (1996), Bobaljik and Thráinsson (1998)], as well as nominal ones [e.g., Cheng and Sybesma (1999), Willim (2000), Trenkic (2004)]. Similar arguments apply at the level of features; that is, although there is a universal set of grammatical features, not all features are instantiated in every language [e.g., Ritter (1993), Chomsky (1998, 2001)]. Furthermore, languages vary in terms of feature strength, with a range of syntactic consequences [Pollock 1989]. Hence, in those cases where the L1 and L2 differ in terms of categories, features, or feature strength, successful L2 acquisition must involve the acquisition of “new” categories and their associated features and feature strength. The interlanguage lexicon, in other words, will include categories and features not found in the L1 lexicon. In contrast, if interlanguage...
3. Lexical categories in L2 acquisition

There has been relatively little discussion of the L2 acquisition of lexical categories as such. Dietrich (1990) and Klein and Perdue (1997) report that nouns, verbs, adjectives, adverbs, and prepositions are present from the beginning. Presumably the absence of detailed discussion of the acquisition of lexical categories reflects the fact that L2 lexical category acquisition has not proved particularly problematic. The L1 grammar includes categories like noun and verb; thus, it would seem uncontroversial to propose that L2 learners start out with the ability to categorize (based on the L1), and that they categorize L2 words accordingly.

However, a different perspective is presented by Lakshmanan (2000) who argues that, in the very early stages of child L2 acquisition, the lexical category verb is not projected\(^1\). Instead, she claims that a combined lexico-functional head is projected. This claim is based on the finding of utterances lacking lexical verbs in child L2, such as those shown in (1) [from Lakshmanan (1993/1994)].

(1) a. This is the boy for the cookies.
   b. The boy with the milk.
   c. This cat and car.

Lakshmanan (2000) suggests that elements like for, with and and are lexico-functional heads with the categorial features [–N, –V] and with Case features. Projection of this “mixed” category is claimed to be a universal stage, regardless of the L2 learner’s L1.

In contrast, as will be discussed below, most other researchers assume either that the initial interlanguage grammar contains distinct lexical and functional categories or that it initially includes only lexical categories. In either case, the L1 is assumed to be crucially involved, at least as far as lexical categories are concerned. While it is difficult to tell, directly, whether lexical categories are transferred as such from the L1 grammar, there are indications that the L1 is implicated. In initial stages, L2 learners transfer properties associated with lexical categories, in particular, headedness. This is true of child and adult L2 acquisition [du Plessis et al. (1987), Hulk (1991) Schwartz and Sprouse (1994), Vainikka and Young-Scholten (1994), Haznedar (1997), van de Craats Corver and van Hout (2000)].

For example, Haznedar (1997) shows evidence of L1 headedness in the early acquisition of English by a Turkish-speaking child, Erdem, who was initially interviewed at the age of 4, after 3 months in the UK. Turkish VPs are head-final, whereas English VPs are head-initial. For the first 3 months of observation, Erdem consistently (almost

---

\(^1\) It is not clear whether this is seen as a problem in syntactic categorization as a whole; Lakshmanan clearly assumes that nouns are identified and that they project NPs.
100%) produced head-final word order, suggesting a transfer of Turkish headedness. For example, he would produce OV utterances like (2a). In the fourth month, VO utterances predominate, as in (2b), suggesting a switch to head-initial order.

(2) a. I something eating
   b. You eating apple

L1 headedness of lexical categories is also found in adult L2 acquisition, although this was originally denied by some researchers [e.g., Clahsen and Muysken (1986)]. Vainikka and Young-Scholten (1994, 1996) examine the L2 German of Turkish and Korean speakers, as well as Romance speakers. VPs in German are head-final, though this is obscured in main clauses because the verb moves into second position (V2). VPs are also head-final in Korean and Turkish. Portuguese and Italian, on the other hand, have head-initial VPs. Vainikka and Young-Scholten (1994) found clear differences between the learners whose L1s were head-final and those whose L1 was head-initial. Over 95% of VPs produced by the three least advanced Turkish/Korean speakers in their study were head-final, as in (3). (Such sentences are ungrammatical in German, because main clauses must have a finite verb in second position.)

(3) a. Oya Zigarette trinken
    Oya cigarette drink-INF
    ‘Oya smokes cigarettes’
   b. Eine Katze Fisch alle essen
    a cat fish all eat-INF
    ‘A cat ate the entire fish’

Four Romance speakers at a similar stage of development show something quite different. In particular, their VPs are head-initial, as shown in (4).

(4) a. Trinke de orange oder?
    drink the orange or?
    ‘(She’s) drinking the orange (juice), right?’
   b. De esse de fis
    she eat the fish
    ‘She’s eating the fish’

In other words, low-proficiency learners of different L1s adopt different word orders in the early interlanguage grammar, suggesting that lexical categories in interlanguage grammars take on properties of lexical categories in the L1.

Hulk (1991) reports on the L2 acquisition of French (head-initial) by speakers of Dutch (head-final) in high school and at university. Using a grammaticality judgment task, she finds considerable evidence, in initial stages, of acceptance of word orders that show L1 headedness (and are ungrammatical in the L2), followed, in later stages, by acceptance of L2 word order and rejection of L1 order. These results suggest resetting from L1 to L2 headedness over the course of development.
To summarize, after an initial period of adopting L1 headedness, resetting to L2 headedness of lexical categories appears to be relatively straightforward. Although L2 acquisition of German word order can be problematic, this appears to relate to the V2 phenomenon, rather than headedness as such.

4. Functional categories in acquisition: Issues of evidence

The status of functional categories in interlanguage grammars is more controversial. There is considerable disagreement over whether or not functional categories are initially realized in interlanguage grammars and over the extent of L1 influence at various points in development. This controversy reflects, at least in part, a disagreement as to what counts as evidence for the presence of a functional category, in particular whether the relevant evidence is primarily morphological or primarily syntactic.

In the so-called morpheme order studies of the 1970s [e.g., Dulay and Burt (1974)], inflectional morphology and function words in L2 were treated as a set of phenomena to be investigated together, although with no clear conception at that time of what the connections might be; supplying morphology in 90% of obligatory contexts was taken to indicate acquisition, following Brown’s (1973) proposals for L1 acquisition. At that time, inflectional morphology was considered independently of other properties of the grammar. Since then, researchers have become interested in the morphology/syntax interface, namely the relationship between syntax, abstract morphosyntax, and overt morphology, particularly with respect to the question of what counts as evidence for the acquisition of functional categories.

In the debate over the status of functional properties in L2 grammars, one can distinguish two main positions: morphology-before-syntax and syntax-before-morphology [see White (2003), Cha. 6]). According to the former view, a very close connection is assumed between overt morphology and abstract syntax, such that the absence of overt morphology is taken to mean the absence of the corresponding functional categories. Furthermore, the acquisition of morphology is assumed to drive the acquisition of syntactic categories. When morphology emerges (or when it is supplied in a certain arbitrary percentage of obligatory contexts), categories are assumed to be in place.

The small clause hypothesis/weak continuity hypothesis for L1 acquisition provides an example of the morphology-before-syntax position. On this view, while functional categories are available in the UG inventory, functional structure in L1 acquisition is initially totally absent [Radford (1990)] or severely limited to one underspecified functional projection [Clahsen (1990/1991)]. A functional category is deemed not to have been acquired until overt morphology associated with that category is consistently produced. In other words, the absence of inflectional morphology or function words in a sentence like Mummy eat cookie (meaning Mummy eats a cookie or Mummy is eating a cookie) implies the absence of the corresponding functional structure in the child’s grammar; such a sentence would be represented as in (5). On this view, then, the initial grammar of the L1 acquirer contains only lexical categories and their projections (NP,
VP, etc.). Functional categories like Infl, Comp, and Det and their projections (IP, CP, and DP) emerge gradually, added as a result of exposure to input, maturation, or both.

Another close link between morphology and syntax that is assumed in this kind of account relates richness of morphology to the feature strength of functional categories. Crosslinguistically, rich inflection seems to go with verb raising [e.g., Vikner (1997), Rohrbacher (1999)]. In other words, overt morphology and syntactic properties go hand in hand. Taking adverbs as marking the left edge of the VP, a main verb to the right of the adverb indicates that it has not raised, as in English (6a), whereas a main verb to the left of the adverb indicates that it has raised, as in French and German (6b) and (6c).

(6) a. Mary often drinks coffee.
    b. Marie boit souvent du café.
    c. Maria trinkt oft Kaffee.

In English, impoverished verbal agreement (only third person singular being marked) is found and lexical verbs remain in VP; in French, verbal inflection is “rich” and lexical verbs raise to I; in German, verbal inflection is again rich and verbs raise to I and on to C. Differences in verb placement are attributed to the presence of strong or weak features; strong features motivate movement, for feature-checking purposes.

Based on examples like these, there seems to be some justification in positing a close connection between overt morphology and the properties of corresponding underlying syntactic categories. Indeed, some researchers argue that L1 acquisition reflects this relationship: the acquisition of verb raising in German L1, for example, is claimed to be contingent on the acquisition of inflection [Clahsen (1988), Meisel (1991)]. Researchers such as Vainikka and Young-Scholten (1994, 1996) have proposed a similar relationship for L2, as we shall see.

However, the hypothesis that there is a direct one-to-one correspondence between overt morphological forms and abstract morphosyntactic representations is misconceived on grounds of descriptive adequacy [Hornstein and Lightfoot (1994), Sprouse (1998), Lardiere (2000), Bobaljik (2003)], since there are languages without rich morphological paradigms that have syntactic properties reflecting strong feature values.

---

2 There is some disagreement in the literature on V2 languages as to whether or not subject-initial main clauses involve V2. If they do, the verb has raised to C; if they do not, the verb has raised to I [Travis (1984), Zwart (1993)].
Afrikaans, for example, has no verbal agreement; nevertheless, the finite verb in main clauses raises to I and then to C. Consequently, children acquiring Afrikaans as their L1 cannot use morphology as the trigger for verb raising. Furthermore, in L1 acquisition of French, a language with relatively rich verbal morphology, verb raising is found in advance of overt verbal morphology [Verrips and Weissenborn (1992)]. Clearly, if there are languages of the world where morphology does not drive syntax, as well as circumstances in L1 acquisition where morphology does not drive the acquisition of syntax, it cannot be the case that there is a strong implicational connection between the two.

Hence, the alternative view has arisen: syntax-before-morphology, according to which absence of morphology is not taken to mean the absence of the corresponding categories, either in acquisition or in the grammars of adult native speakers; rather, syntactic evidence must be taken into consideration in determining the status of functional categories. Along these lines, many accounts of L1 acquisition assume full competence. In other words, the initial L1 grammar includes abstract functional categories from the beginning. Syntax is prior to morphology; the abstract skeleton is there even though it is not necessarily filled in with overt morphological material [e.g., Hyams (1992), Poeppel and Wexler (1993), Borer and Rohrbacher (2002)]. On this kind of account, the sentence *Mummy eat cookie* would have a representation as in (7), even though inflectional morphology and function words have not been produced.

(7)

If underlying categories and surface forms are distinct, such that abstract categories and features are represented in the grammar in the absence of any corresponding overt morphology, then it is inappropriate to concentrate on overt morphology as the sole means of determining the presence or absence of syntactic categories in interlanguage grammars. In other words, there is no reason to expect that overt morphological paradigms drive the acquisition of syntactic categories in L2; one needs to look at L2 syntax, as well as (or instead of) L2 morphology, in order to determine the status of functional categories in the interlanguage grammar.
5. Functional categories in the L2 initial state and in L2 development

In L2 acquisition, two opposing positions about the status of functional categories in the early grammar have been adopted, corresponding to the morphology-before-syntax approach and the syntax-before-morphology approach [see White (2003) for discussion]; each one is highly dependent on assumptions about the relevant evidence, particularly the significance attributed to the absence of overt morphology (or failure to supply such morphology consistently).

5.1. Morphology-before-syntax

According to the Minimal Trees Hypothesis [Vainikka and Young-Scholten (1994, 1996), see also Eubank (1996)] or the modulated Structure Building Hypothesis of Hawkins [(2001), see also Wakabayashi (1997)], the initial interlanguage grammar is a grammar containing lexical categories (drawn from the L1 grammar) but no functional categories. This claim is motivated by the absence of consistent use (defined by Vainikka and Young-Scholten as less than 60%) of inflectional morphology and function words in the early stages of L2 production. In other words, spontaneous production data are taken to provide a relatively direct and reliable window onto the underlying grammar: if certain forms (function words and inflectional morphology) are absent in production, the underlying categories associated with them are assumed to be lacking as well.

Vainikka and Young-Scholten (1994) point out that the least advanced Turkish/Korean speakers and the least advanced Romance speakers learning German as an L2, as discussed above, show the following characteristics in their production: (a) correct subject–verb agreement is low; instead, infinitives or bare stems or default suffixes predominate over correctly inflected finite verbs; (b) modals and auxiliaries are very rare; (c) there is no evidence that the verb raises out of the VP; (d) there are no wh-questions or subordinate clauses introduced by complementizers. These characteristics suggest to them that IP and CP are lacking. Thus, an utterance like (3a) would be represented as in (8), parallel to equivalent representations postulated for L1 acquisition [as in (5)].

\[
\text{(8)}
\]

Variable production (morphology showing up in fewer than 60% of obligatory contexts) or absence of production altogether is taken to reflect a property of the underlying grammar. At this stage, it is claimed that functional categories and features are not represented at all in the interlanguage grammar.

While the Minimal Trees Hypothesis and the Structure Building Hypothesis share the assumption that the initial interlanguage grammar lacks functional categories, they
differ radically in their claims concerning what happens to these categories in the course of L2 acquisition. According to Vainikka and Young-Scholten, functional categories appropriate to the L2 are added to the interlanguage grammar during the course of acquisition, there being no L1 influence in this domain. Categories emerge gradually, triggered by L2 input, and are added to the representation from the bottom up (i.e., CP could not be acquired before IP). In contrast, Hawkins (2001), starting from the same assumption, namely that only lexical categories are represented in early grammars, claims that once functional categories/features are added, these consist of the categories and features found in the L1 grammar. Furthermore, L2 learners are claimed to be unable to acquire new (uninterpretable) features of functional categories. In other words, post-puberty L2 learners are “stuck” with those functional categories and features realized in the L1 grammar [Tsimpli and Roussou (1991), Hawkins and Chan (1997), Hawkins (2003), Tsimpli (2003)]. This position is known as the Failed Functional Features Hypothesis (FFFH), the Representational Deficit Hypothesis or the Impaired Representation Hypothesis.

5.2. Syntax-before-morphology

In contrast, a number of researchers argue in favor of the syntax-before-morphology position for L2, maintaining that the crucial issue is not whether L2 learners consistently get the surface morphology right but whether abstract morphosyntactic features are represented in the interlanguage grammar, with their associated syntactic consequences [Schwartz (1991), Epstein, Flynn and Martohardjono (1996), Haznedar and Schwartz (1997), Lardiere (1998a, b), Prévost and White (2000), Ionin and Wexler (2002), see White (2003) for an overview]. In other words, these are full competence theories of L2: there is no initial purely lexical stage; rather, functional categories are present from the beginning, possibly transferred from the L1, in which case changes may have to be made with regard to which categories and features are represented in the longer term, since categories and features found in the L1 may prove to be inappropriate or inadequate for representing the L2.

A number of studies have examined production data from child and adult L2 learners, looking for evidence of abstract syntactic knowledge associated with morphosyntactic features. Many studies focusing on L2 English show use of tense and agreement morphology to be lower than 50% (often considerably lower). This is true of child [White (1992), Haznedar (2001), Ionin and Wexler (2002)] and adult L2 learners [Lardiere (1998a, b)], and of the initial [Haznedar (2001)] and end states [Lardiere (1998a, b)]. At the same time, these authors report other characteristics of production

3 Wakabayashi (1997) offers a position between these two. According to him, some but not all features will transfer. This will depend on whether properties of the L2 input motivate an L1-based analysis.

4 In the case of L2 French and German, use of tense and agreement morphology is reported by many to be much higher than is the case for L2 English [e.g., Grondin and White (1996), Parodi (2000), Prévost and White (2000)].
data that implicate abstract categories and features. L2 learners of English, for example, are sensitive to a number of syntactic properties associated with the functional category Infl, including the presence of overt subjects, suggesting that they raise to Spec IP to check their features, and nominative subject pronouns, suggesting case checking in Infl. Furthermore, CPs are clearly present (as shown by the presence of embedded clauses and by the incidence of wh-questions and of inversion in yes-no questions), even in the absence of consistent inflectional morphology [Gavruseva and Lardiere (1996), Haznedar (2003)].

To summarize so far, in L2 acquisition a divergence is often found between surface inflection and more abstract syntactic properties associated with morphosyntactic features and feature strength. L2 learners of various languages are relatively inconsistent, though by no means random, in their use of certain kinds of morphology; at the same time, they are highly accurate on related syntactic properties. These characteristics appear to be true of initial and end-state learners, as well as learners whose interlanguage grammars are undergoing development. Such results are inconsistent with the morphology-before-syntax hypothesis, since their syntactic knowledge does not depend on overt morphology in any way.

An additional debate concerns whether learners are “stuck” with functional categories and features available in the L1, as claimed by proponents of the FFFH. Hawkins and Chan (1997) investigate the syntactic consequences of the purported absence of a [±wh] feature in C in the interlanguage grammar of Chinese-speaking learners of English. Hawkins and colleagues extended the FFFH to abstract features in T [±past] and in D (gender), claiming that adult learners’ failure to consistently supply overt tense morphology or gender agreement reflects the absence of equivalent features in the L1 grammar. Thus, Mandarin-speaking learners of English are claimed to lack a [±past] feature in the interlanguage grammar [Hawkins (2000), Hawkins and Liszka (2003)], while English-speaking learners of Romance languages lack gender [Franceschina (2001), Hawkins and Franceschina (2004)]. In a related vein, Tsimpli (2003) argues that L2 learners have problems with the feature [±definite] if this is not represented in the L1. Once again, considerable significance is attributed to the absence or inconsistent use of overt morphology, which is taken to indicate a lack of features, but not necessarily the absence of functional categories as such.

The debate over whether or not L2 learners are stuck with L1 categories or features is ongoing—see Hawkins (2001) and White (2003) for overviews from each side of the debate. On the whole (pace Hawkins and colleagues), the research is consistent with the hypothesis that L2 learners can add new categories and features to the interlanguage lexicon, resulting in L2-appropriate syntax, as will be discussed in the next section.

6. Acquiring versus losing categories and features

A number of researchers have pointed out that an implication of the Minimalist program for L2 acquisition/bilingualism is that there is only one computational system but two
lexicons [Dekydtspotter, Sprouse and Anderson (1997), MacSwan (2000), van de Craats et al. (2000)]. The relationship between the two lexicons is therefore crucial. Robertson and Sorace (1999) suggest that, initially, L1 lexical entries, including abstract functional features but excluding phonetic content, are copied and used as templates for forming entries in the L2 lexicon [see van de Craats et al. (2000) for related proposals]. These L1-based lexical entries are, in many cases, revised in the light of the L2 input.

Category acquisition, then, involves two things, at least potentially: (i) the addition of new categories (and features) to the interlanguage lexicon (i.e., categories not present in the L1 but motivated in the L2); (ii) the loss of categories (and features) from the interlanguage lexicon (i.e., categories/features which were copied from the L1 lexicon and which are not in fact required in the L2). So far, there has been considerable research investigating the acquisition of new categories/features required to appropriately represent the L2. Indeed, recent debate has centered on whether or not this is possible, including much of the research described above.

Let us consider a specific example to illustrate the acquisition of new L2 categories. Many researchers have extended to all languages as Pollock’s (1989) proposal that both French and English have a split IP; in other words, this is taken to constitute a universal. Arguing against this position, Thráinsson (1996) and Bobaljik and Thráinsson (1998) propose a split-IP parameter, such that languages which have thematic verb raising in non-V2 environments\(^5\) (for example, Icelandic) have two distinct functional categories, Agr and T, projecting both an AgrP and a TP [(see (9a)]. In contrast, languages which lack verb raising (for example, English) have only one category, I, which projects an IP, as in (9b). [In a related vein, Iatridou (1990) argues that there is no reason to postulate an Agr projection in English, in contrast to French.]

\(\text{(9)}\)

\[
\begin{align*}
(a) &\quad \text{AgrP} \\
&\quad \text{Agr} \\
&\quad \text{TP} \\
&\quad \text{T} \\
&\quad \text{VP} \\
&\quad \text{V} \\
(b) &\quad \text{IP} \\
&\quad \text{I} \\
&\quad \text{VP} \\
&\quad \text{V}
\end{align*}
\]

Along with verb raising, [+split-IP] languages allow object shift (10a) and transitive expletive constructions (10c), whereas [–split-IP] languages do not (10b and 10d) [examples from Bobaljik and Thráinsson (1998)].

\(\text{(10)}\)

(a) Ég las Þrjár bækur ekki.
   I read three books not
(b) *I did three books not read.

\(^5\) V2 is the result of further verb raising to C and is subject to a different parameter; see below.
c. Það hefur einhver köttur étid mysnar.
   there has some cat eaten mice-the

d. *There has a cat eaten the mice.

As far as morphology is concerned, only split-IP languages allow a verb to carry separate tense and agreement markers at the same time (each checked in a different head), as shown for Icelandic in (11a), in contrast to English (11b).

(11) a. kasta-ðí-r
   throw-past-2s

   b. *trembled-s

Conradie (2002) argues that Afrikaans is a split-IP language. Even though Afrikaans lacks a morphologically distinct realization of tense and agreement (since it lacks agreement morphology), it shows all the syntactic effects of the split-IP setting, including verb raising in non-V2 contexts (12a), object shift (12b), and transitive expletives (12c) [examples from Conradie (2002)].

(12) a. Helgi lees dikwels boeke.
   Helgi reads often books

   b. Ek het nie drie boeke gelees nie.
   I have not three books read not

   c. Daar het ‘n kat die muise geëet.
   there has a cat the mice eaten

As far as L2 acquisition is concerned, using a variety of tasks (sentence manipulation, grammaticality judgments, and truth value judgments), Conradie demonstrates that advanced proficiency English-speaking learners of Afrikaans generally perform like native speakers in accepting or producing transitive expletives and object shift constructions, as well as recognizing subtle scope effects in the latter case. In other words, L2 learners can switch from a grammar with a single category I to a grammar with distinct Agr and T categories, suggesting that new categories can be added to the interlanguage lexicon.

The opposite situation has been discussed much less, that is, whether it is in fact possible to “lose” categories or features required in the L1 but not needed to represent the L2. If a functional category (or feature) is present in the L1 grammar and not required in the L2, is this category lost or is it in fact represented in the interlanguage grammar with certain visible consequences? Could it be harder to lose categories than to add them?

Of course, in many cases, it will be impossible to tell. For example, if the L1 has a [±past] feature, say, and the L2 does not, it is not clear what kind of error would arise by assuming this feature in the L2. Nevertheless, there are cases where the effects of a failure to lose categories may be visible. In particular, there are some cases of residual errors in advanced speakers, which may reflect the fact that certain categories get copied into – and not subsequently deleted from – the L2 lexicon [Robertson and Sorace (1999)].
In this context, let us once again consider the split-IP parameter. A well-known error found in the L2 acquisition of English by speakers of French involves the raising of thematic verbs past adverbs [White (1990/1991, 1991, 1992)], as in (13a). This is a common and persistent error, found not only in the initial stages but also in the utterances of advanced L2 speakers. At the same time, L2 learners do not make errors with negative placement; forms like (13b) are rare or non-existent.

(13) a. Mary drinks often coffee.
   b. *Mary drinks not coffee.

One possible account of this error is that French-speaking learners of English inappropriately assume two categories, T and Agr, on the basis of the L1 grammar [as in (9a)] and fail to replace them by a single category, I. It is only if there are two categories above the VP that one can account for the differential treatment of verb raising in the context of adverbs and negation. Assuming that adverbs are represented lower than negation, between T and VP, while negation occurs between Agr and TP, the existence of errors like (13a) and the absence of errors like (13b) suggests that the verb raises to T (the lower functional category), over the adverb, but not to Agr (the higher one). In contrast, if L2 learners were representing English sentences with a single category, I, with both adverbs and negation located between I and VP, they would be expected to make parallel errors, raising the verb to I over both adverbs and negation. In other words, errors like (13a) and (13b) would be expected to co-occur, or correct word order without raising in either case would be expected. The persistence of errors like (13a) and the absence of errors like (13b), then, argues for two distinct functional categories in the interlanguage grammar, in other words, failure to “lose” a category.

Turning to another example, namely V2, a similar situation has been reported: difficulty in losing features/feature strength. V2 is driven by a strong feature in C, which is present in the grammars of German and Afrikaans but absent in English; it has been shown that L2 learners can acquire this feature or reset its strength. For example, in addition to showing that L2 learners of Afrikaans reset the split-IP parameter, Conradie (2002) shows that they reset the V2 parameter, accepting V2, as in (14a) and rejecting V3, as in (14c).

(14) a. Die jakkals het Olaf met hierdie geweer geskiet.
   the fox has Olaf with this gun shot
   ‘Olaf has shot the fox with this gun.’
   b. *The fox has Olaf shot with this gun.
   c. *Die jakkals Olaf het met hierdie geweer geskiet.
   d. The fox Olaf has shot with this gun.

Similar findings have been reported by other researchers for L2 German, often after an initial stage of L1 transfer [e.g., du Plessis et al. (1987)]7. In other words, it is possible for

---

6 Previous accounts of the persistence of verb raising have been couched in terms of optional verb raising, unspecified feature strength, etc. [White (1992), Eubank (1996)].

7 But see Clahsen and Muysken (1986) for a contrary view.
learners to acquire a strong feature value in the L2 when the L1 has the corresponding weak value.

Again, the reverse situation is less frequently discussed but Robertson and Sorace (1999) offer relevant speculation. They identify residual V2 effects in the L2 grammars of advanced German-speaking learners of English. In particular, along with producing appropriate V3 word orders, as in (14d), a minority also occasionally produce/accept V2 orders as well, as in (14b). According to Robertson and Sorace, such errors reflect the fact that, where an L1-based lexical item has been copied into the L2 lexicon but is not, in fact, required for the L2, it may never be fully expunged from the L2 lexicon. In this case, the interlanguage lexicon of German-speaking learners of English includes an entry [+strong C], required for German but not for English. Occasionally, this strong feature is selected when an L2 sentence is produced, resulting in residual V2 effects.

Other potential examples of failure to lose L1 categories concern properties of DPs. For example, Leung (2002) explores the L2 acquisition of English and French by speakers of Cantonese and Vietnamese. Cantonese and Vietnamese lack the category D; instead they have classifiers (CL) [Cheng and Sybesma (1999)]. Thus, L2 learners of English or French with these L1s must acquire D and lose CL. Classifiers differ from determiners in that they can co-occur with demonstratives and possessives, as shown in the Cantonese example in (15) [from Leung (2002)].

(15) a. Keoi3 go3 mui2 hou2 leng3
   his/her CL sister very beautiful

   b. *His the sister is very beautiful

Thus, if the category CL is maintained while the category D is being acquired, sentences like (15b) are predicted to be acceptable. Leung’s results, indeed, suggest that CL is a category represented in the interlanguage grammars of acquirers of L2 English (Cantonese-speaking subjects) and L2 French (Vietnamese-speaking subjects); sentences like (15b) are accepted to a considerable extent, especially by beginners, who at the same time show other behavior indicative of the presence of determiners. In other words, CL and D are found together. With more advanced proficiency, the category CL is dropped, or at least forms like (15) are no longer accepted, suggesting that this category can be lost.

Another persistent error which is amenable to an analysis involving the maintaining of L1 categories or features not required by the L2 concerns the feature gender. There have been many studies on the L2 acquisition of gender, with debate over the extent to which it is possible to acquire gender in the L2 if the L1 lacks gender [see, for example, Franceschina (2001) and Hawkins and Franceschina (2004) versus White et al. (2004)]. In contrast, there has been little consideration of whether there are problems in losing grammatical gender. Nevertheless, a well-known and persistent L2 error is amenable to such an analysis. French is a language with gender agreement; English lacks grammatical gender. French-speaking learners of English appear to impose gender agreement on English nouns, as would be required in the L1, for example, as in (16).

(16) John sat on her chair (meaning his chair)
In French, the noun *chaise* (‘chair’) is feminine and there is gender agreement between nouns and determiners. Errors like (16), then, suggest difficulty in losing gender features, which are imposed on L2 nouns even though they are not required.

7. Discussion

One might ask why it is that L1 categories and features are sometimes found only in the early stages of L2 acquisition but at other times persist even in the grammars of advanced L2 speakers. One possibility is that learnability considerations come into play. While there will be positive evidence to motivate the addition of new categories, it is not clear what evidence would motivate the loss of L1 categories, unless parameter resetting (or other forms of mutual exclusivity) are involved.

Consider once again the split-IP parameter. Here, according to Bobaljik and Thránsson (1998), there is a clear diagnostic for each setting of the parameter, namely [± verb raising]. In that case, a single I rather than a split T and Agr should be motivated for the grammars of French-speaking learners of English, on the basis of sentences like (6a), repeated here as (17). In such cases, the verb has not raised, and the setting is [−split-IP].

(17) Mary often drinks coffee

As we have seen, French-speaking learners appear to maintain the [+split-IP] value of this parameter, despite the presence of positive evidence motivating resetting.

However, it may be that, in this case, another aspect of the English input misleads the L2 learner regarding verb raising. In English, nonthematic verbs (such as the copula or auxiliaries) raise, as shown in (18); this is also the case in French.

(18) a. Mary is often busy.
    b. Mary has often drunk too much coffee.

One possibility, then, is that French-speaking learners of English are misled by the possibility of raising nonthematic verbs in both languages into maintaining the [+split-IP] value of the parameter, hence continuing to represent the L2 with two verbal functional categories rather than one. Vainikka and Young-Scholten (1996), in fact, propose something similar, namely that all learners of English, regardless of L1, will be misled by the raising of nonthematic verbs in English. However, thematic verb-raising errors are not found if the L1 is a language like Russian or Mandarin, which lacks verb raising [Lardiere (1999), Ionin and Wexler (2002)]. Furthermore, Yuan (2001) shows that if Mandarin is the L2, French-speaking learners do not assume verb raising in the L2. He points out that Mandarin has no verb raising at all, either of thematic or nonthematic verbs. In our terms, in this case, French speakers are able to reset from [+split-Infl] to [−split-Infl] early on; they have no difficulty in losing the L1 categories. Putting these observations together, properties of the L2 input can indeed mislead certain learners into maintaining L1-based categories, which are not in fact required in the L2.
French-speakers are misled when they learn English (where auxiliary verbs raise) but not when they learn Mandarin (where they do not).

In conclusion, it has been suggested in this chapter that the relevant evidence for demonstrating category acquisition is not necessarily morphological. When syntactic evidence is taken into consideration, it is clear that non-native speakers can successfully acquire L2 syntactic categories, both lexical and functional. A number of familiar L2 errors have been reconsidered here in terms of the ability (or lack thereof) to lose L1 categories. Losing categories or features, which have been copied from the L1 lexicon is not always a straightforward matter and can result in subtle and lasting effects on the interlanguage grammar, including residual optionality, as Robertson and Sorace (1999) suggest.

References


Radford, A. (1990), Syntactic Theory and the Acquisition of English Syntax (Blackwell, Oxford).


This page intentionally left blank
Chapter 24

THE DEVELOPMENT OF CATEGORIES IN THE LINGUISTIC AND NONLINGUISTIC DOMAINS: THE SAME OR DIFFERENT?

DIANE POULIN-DUBOIS

Concordia University
In cognitive science, the literature on categorization rarely references the developmental research on concept formation, although the reverse is not necessarily true. Fortunately, a bridge between these “two solitudes” has been built over the last decade, as shown by the inclusion of a few chapters on development in a recent “state-of-the-art” book on the topic [Murphy (2002)]. The content of the present handbook confirms the relevance of children’s categorization abilities to cognitive science, with the inclusion of five chapters dedicated to developmental issues. One major debate on which a developmental approach can shed a unique light is whether categorization is highly determined because there is a natural discontinuity in the world that simply needs to be detected or whether categories are derived from language. Thus, it becomes important to understand the category-forming abilities that preverbal infants possess as they are at the origins of adults’ complex categories.

In her provocative chapter, Rushen Shi proposes that infants form broad, language-universal syntactic categories that are equivalent to superordinate taxonomic categories. These basic, foundational categories correspond to function and content words. This fundamental dichotomy is assumed to play a role in language acquisition in a number of ways: by making the formation of more refined grammatical categories easier, by facilitating word segmentation, and by constraining the possible mapping of meanings to words. Shi argues that speech input provided to infants contains acoustic, phonetic, and phonological cues to the language-universal distinction between function and content words. This proposal contrasts with research based on expressive language that shows that, in contrast to content words, function words are extremely rare and only emerge gradually in language development. For instance, function words account for only 1–4% of the first 50 words produced by English- and Italian-speaking children, and only 7–8% of the first 200 words understood [Caselli, Casadio and Bates (2001)]. It has also been shown that the early use of function words does not predict the development of grammatical skills [Bates et al. (1994)]. Nevertheless, it is worth noting that research based on analyses of naturalistic productions suggests that function words might be underrepresented in parental reports (such as the MacArthur Communicative Development Inventories) and develop earlier and faster in languages such as French [Bassano, Mallochon and Eme (1998)].

In support of her proposal, Shi reviews recent experimental research, including her own, that suggests that function and content words have distinct acoustic, phonological, and distributional properties. More specifically, she shows that function words are minimized across languages (e.g., shortened vowel duration, smaller number of syllables) and are more frequent in input than content words. Interestingly, the overlap between these two categories of words for any given cue suggests that a correlation among multiple cues is necessary for categorization. Although Shi cites neural network simulation studies that can assign novel words to these categories on the basis of such cues, there is also a body of experimental work supporting the hypothesis that infants are sensitive to these cues. Surprisingly, using a habituation procedure, newborns appear to detect a category change between function and content words, manifested by a change in their sucking rate. Such a precocious ability at birth seems at odds with research on visual categorization that shows that infants are not able to explicitly categorize complex
stimuli on the basis of correlations among stimulus features until they are 9 or 10 months of age [Rakison and Poulin-Dubois (2002), Younger (2003)]. In contrast, Shi reports the results of one of her studies in which the recognition of function words (as opposed to nonsense words) developed between 8 and 13 months. Interestingly, this coincides with the developmental changes in the visual domain, and therefore might be a more valid measure of function versus content word classification [Shi, Werker and Cutler (2003, 2004)]. To what extent these broad categories assist infants in language learning tasks remains to be determined. Infants as young as 7.5 months are able to complete word segmentation [Jusczyk (2003)]. Nevertheless, at this age they still show poor representation of function and content words. One direct way to test the link between classification of word classes and grammatical development would be to conduct longitudinal studies. Certainly by the end of the first year, infants can already make fine discriminations within each of the two grammatical word categories. For example, infants as young as 13 months can discriminate among items within the larger content word category, shifting their attention to different attributes of objects in a category formation task as a function of the type of words presented (e.g., adjectives or nouns) [Waxman (2002)]. Similarly, it has been shown that infants can use an indefinite article to infer whether a word is a proper or common noun by the age of 17 months [Katz, Baker and Macnamara (1974)].

The idea of a superordinate-to-basic shift in the development of phonetic categories is consistent with recent research on the development of object categories. In the mid- to late 1970s Rosch argued that there is a basic-to-superordinate shift in the development of natural categories. This hypothesis, widely accepted until recently, was based on data showing that categorization at the basic level is predominant in children and adults, and that basic-level terms (e.g., dog) emerge in children’s vocabulary before superordinate terms (e.g., animal). However, the advent of methodologies suitable for testing categorization in preverbal infants has generated a flurry of cross-sectional and longitudinal studies that have demonstrated that category learning in the object domain (e.g., natural and artifactual) may proceed from broad to narrow (or global to basic) during the first 2 years of life [Mandler, Bauer and McDonough (1991), Poulin-Dubois, Graham and Sippola (1995), Mandler and McDonough (1998), Mandler (2000), Pauen (2002)]. For example, studies that have used the object-examination task, a procedure appropriate for young infants, have yielded results indicating that infants as young as 7 months can categorize objects at the global level, such as animals and vehicles [Mandler and McDonough (1993, 1998), Oakes, Coppage and Dingel (1997)]. By only a few months later, infants can make finer discriminations between basic-level categories, such as between dogs and cats or cars and trucks. This suggests that the broad categories of animals and vehicles only resemble superordinate domains in that they do not possess a hierarchical structure that would include basic-level categories. The cognitive mechanisms by which these global categories are formed are the subject of much controversy [see Quinn (2002) for a review].

Perception of certain speech contrasts is said to be categorical if the sounds belonging to different phonetic categories are easily distinguished, whereas sounds belonging
to the same phonemic categories are poorly discriminated. In a classic study, Eimas and his colleagues showed that infants as young as 1 month not only discriminate among consonants that differ in voice onset time but also show, like adults, no ability to distinguish comparable voicing differences between different instances of the same consonants [Eimas et al. (1971)]. Based on these findings, it was argued that the origin of categorical perception of speech contrasts is innately determined. It is a generally accepted fact that categorical perception is not unique to speech processing in either infants or adults. The chapter by Martijn Goudbeek and his colleagues concerns the acquisition of phonetic categories in adults as it compares with infants’ ability to form such categories without training. Given that infants show extreme sensitivity to the statistical properties of speech signals, the authors test the hypothesis that adults form speech and nonspeech categories that are defined in a two-dimensional psychophysical space. Their findings indicated that adults can form categories based on one dimension; however, they found it extremely difficult to form both types of categories when defined over two dimensions, even with feedback. This pattern of results bears a striking resemblance to the loss of nonnative phonetic contrast discrimination observed in adults and children after 10–12 months of age. Many interpretations of the different findings for infants and adults are discussed, including the amount of exposure, the effect of explicit feedback, and the interference of phonetic categories from L1 on L2 learning in adults. The stimuli used by Goudbeek and his colleagues in their speech category experiments were synthesized versions of Dutch front vowels. It is doubtful that infants would have performed better than adults in forming categories of that type, because discrimination of contrasts involving vowels is continuous in both adults and infants [Jusczyk (2003)]. Nevertheless, this research raises a number of important issues that remain largely unexplored in the literature. One of them concerns the common and unique mechanisms by which infants and adults form auditory and visual categories. This research suggests that statistically based category learning is not equivalent in the auditory and visual modality in adults. In the developmental literature, there is evidence that by 9 or 10 months of age, infants show a recategorization ability whereby new category boundaries are superimposed on earlier perceptual categories. The emergence of explicit categorization of complex stimuli on the basis of correlations among stimulus features has been demonstrated to be the common mechanism underlying developmental changes in both speech perception and visual categorization at that age [Lalonde and Werker (1995)]. A comparison of findings in the adult and developmental literature will certainly benefit our understanding of categorization processes.

In her chapter, Marie Labelle considers the development of linguistic categories that are more abstract than phonetic categories, namely grammatical categories. This chapter complements Shi’s nicely because it considers the acquisition of subcategories within the content word category. Some researchers have argued that children’s early utterances are item-based, that is, they are structured around specific words and phrases, not around syntactic categories [Tomasello (2001)]. The kind of abstract syntactic categories used by adults are assumed to emerge only gradually during the preschool period. For example, both experimental and observational studies have shown that
before the age of 3 or 4 years, children use determiners such as *a* and *the* with sets of nouns that overlap only minimally, suggesting that these two determiners are not included in the same category [Pine and Lieven (1997)]. In contrast, it is believed that children form grammatical categories such as concrete nouns by the age of 24 months [Tomasello and Olguin (1993)]. Labelle believes that children do not form categories of words by semantic bootstrapping, and cites the cases of English words that are both nouns and verbs and languages for which this is the norm (e.g., Tagalog). She argues that in order to home in on the adult grammar, children must process the distributional properties of lexical items. More specifically, she proposes that word order, morphological cues, and co-occurrence restrictions are the three types of formal cues that children use to categorize words. With regard to the role of word order, I believe that the fact that infants treat nouns as grammatical categories shortly after they show sensitivity to word order in both expressive and receptive language provides support for the role of this type of cue. However, as the author points out, word order cannot be of much help in explaining the differential prevalence of nouns and verbs across languages. Another cue that might help children to categorize words is the presence of function words and inflectional morphemes. As Labelle notes, one must be able to categorize a grammatical element as marking a grammatical function in order to use it as a cue to categorize a word. However, one might argue that detection of grammatical elements and sensitivity to their distribution might be sufficient to formulate part-of-speech categories. As mentioned before, children still do not seem to classify two determiners in the same class after the age at which they show that they possess a grammatical category of nouns. Labelle concludes that, although there is sufficient support for the hypothesis that children are sensitive to distributional information, many competing proposals exist regarding the distributional learning mechanisms that may account for the acquisition of grammatical categories. Again, there is an interesting concordance in verbal and nonverbal categorization, given that notions such as cue validity are assumed to play a role in both domains. Also, it is unlikely that distributional learning is unconstrained. Rather, a range of constraints has been proposed, which vary from domain-specific innate linguistic knowledge to domain-general statistical learning mechanisms and length of attention span. Clearly, language input contains many cues about grammatical categories but the saliency of these cues varies as a function of the child’s learning abilities.

In her chapter, Eve Clark addresses the contribution of cognitive and social factors to the development of semantic categories. There is currently much interest in the issue of how language and cognition interact to produce early word meanings. In accordance with the Whorfian view that meanings are structured through exposure to language, early research on word meaning development emphasized the role of linguistic input in children’s concept formation [Brown (1958)]. Following Piaget’s theory of the emergence of cognitive skills in preverbal infants, the view that children’s first words map onto concepts became dominant [Nelson (1974)]. Although the assumption that children bring their own concepts to the word-learning task is still very present in the current literature on early word meanings, many researchers have advocated a more
interactive view of how semantic development arises. According to this view, semantic development takes place through an early interaction between nonlinguistic conceptual development and the way that the language input categorizes spatial meanings. The relative contributions of nonlinguistic concepts and language-specific semantic systems are hypothesized to vary along a continuum from linguistic to cognitive dominance [Gentner and Boroditsky (2001)]. At one end of the continuum are words that refer to perceptually individuated items (e.g., concrete nouns), whereas words that cannot exist independently of language are at the linguistic end of the continuum (e.g., determiners and conjunctions). As Clark points out, categories sometimes map directly onto the conceptual categories that children possess, whereas in other cases, the mapping requires the development of new categories. The acquisition of spatial semantic categories provides a good example of the interplay between the universal conceptual abilities that infants develop during the first year of life and language-specific principles of semantic categorization. Clark’s own work and recent findings from crosslinguistic studies show how children construct semantic categories related to space on the basis of how spatial words are used in the input (e.g., frequency, consistency, etc) and also by relying on perceptual sensitivities and conceptual biases. One important contribution of this chapter is to highlight the fact that infants bring not only their conceptual knowledge about the physical world to the word-learning task but also their social-cognitive skills, including the ability to establish joint attention. I would add that infants’ sensitivity to the sociopragmatic cues provided by agents in a verb-learning context is also one part of the infant’s word-learning toolkit [Poulin-Dubois and Forbes (2002) in press]. Another significant contribution of Clark’s paper is the concept of emergent and robust categories. Emergent categories reflect temporary distinctions that children make in their use of language, even when the language they are learning does not support them. For example, words coined by English-speaking children express a distinction between inherent and temporary properties despite the absence of any such distinction in English (e.g., crumby vs. crumbed). In contrast, robust categories are those that are supported in the speech addressed to children, whether by syntactic form classes, morphology, or the lexicon. One question that remains is the fate of emergent categories in the languages of bilingual children who learn two languages that differ in their conventional expression of these categories. Will these emergent categories survive in both languages if support is found in L1 or L2?

The topic of bilingualism and syntactic categories is directly addressed in Lydia White’s chapter. More specifically, her chapter concerns the representation of syntactic categories in L2. Many researchers have proposed that languages differ in terms of the lexical and functional categories that they instantiate. Thus, successful acquisition of L2 might be challenged if L1 and L2 have different categories because L2 acquisition will require the acquisition of categories not found in the L1 lexicon. In the case of lexical categories, White reviews studies that put forward conflicting accounts of L2 acquisition. On the one hand, it has been argued that L2 does not initially contain lexical categories such as verbs (e.g., the boy with milk). In contrast, most researchers assume that the L1 is involved in the formation of L2 lexical categories. White concludes that L1
properties (e.g., headedness) are transferred to L2 in the initial stages of L2 acquisition. Because research with both children and adults is reported, these findings appear to tease apart the effect of fluency in L1 on transfer to L2. I also wonder whether the initial L2 categories have the same status as the emergent categories that Clark discusses in her chapter. In other words, do conceptual and linguistic biases carry the same weight when building lexical categories? With regard to functional categories, there is some controversy with regard to the role that morphology and syntax play in L2 acquisition of these categories. White sides with the syntax-before-morphology proponents, based on the presence of categories such as complementizers in the absence of consistent inflectional morphology. Again, the conditions under which a bilingual speaker will lose L1 categories not required in L2 and acquire new L2 categories remains controversial. It is important to note that the issue of how categories constructed in one domain are “projected” to another domain is also relevant for object categories. For example, as mentioned before, unlike adults and older children, preschool children treat nominal kinds as taxonomic categories, focusing on characteristic instead of defining features. Maybe even more relevant in the present case is the word-learning principle of mutual exclusivity, i.e., one word per referent, that very young monolingual and bilingual children follow when exposed to a novel word [Frank and Poulin-Dubois (2002)]. Interestingly, in the area of semantic development, rigid adherence to this principle would prevent children from maintaining lexical items in L1 and adding new words to their L2 lexicon. But, as they become more proficient in L2, bilingual children are less likely to honor the mutual exclusivity principle, particularly when learning new words in a dual-language context. Paradoxically, a similar mutual exclusivity principle is hypothesized to motivate the loss of L1 syntactic categories not required in L2, such as parameter resetting. In sum, the issue of how categorization skills are transferred across “domains” is a fascinating one, and White shows how syntactic categories might be added or lost through this process.

In conclusion, all five chapters spanning the development of categories in language address important issues that both overlap and differ from those discussed in current research in the nonlinguistic domain. The challenge ahead is to identify the domain-specific and domain-general mechanisms involved in category formation across both domains.

References


PART 5

NEUROSCIENCE OF CATEGORIZATION
AND CATEGORY LEARNING
Chapter 25

MULTIPLE SYSTEMS OF PERCEPTUAL CATEGORY LEARNING: THEORY AND COGNITIVE TESTS

F. GREGORY ASHBY AND VIVIAN V. VALENTIN

University of California, Santa Barbara

Contents

Abstract 548
1. Introduction 549
2. Two Category-Learning Tasks 550
3. COVIS 550
4. The COVIS explicit system 552
   4.1. Switching attention in the explicit system 554
   4.2. Long-term storage of explicit category knowledge 556
5. The COVIS procedural-learning system 557
6. Competition between the COVIS explicit and implicit systems 560
7. Dissociations between rule-based and information-integration category learning 561
8. Conclusions 563
Appendix A 564
   A.1. Network implementation of the explicit system 564
   A.2. Network implementation of the implicit system 566
Acknowledgment 568
References 568
Abstract

The competition between verbal and implicit systems (COVIS) theory of category learning [Ashby et al. (1998)] postulates two systems that compete throughout learning – a frontal-based explicit system that uses logical reasoning and depends on working memory and executive attention, and a basal ganglia-mediated implicit system that uses procedural learning. The implicit system can learn a wide variety of category structures, but it learns in a slow, incremental fashion and is highly dependent on reliable and immediate feedback. In contrast, the explicit, rule-based system can learn a fairly small set of category structures quickly – specifically, those structures that can be learned via an explicit reasoning process. This theory is described in detail and a variety of cognitive-behavioral experiments are reviewed that test some parameter-free \textit{a priori} predictions made by COVIS.
1. Introduction

All animals must learn to categorize objects and events in their environment to survive. Is this mushroom edible or poisonous? Is the object behind the bush a deer or a wolf? Correct classification allows animals to select the appropriate approach or avoidance response to nutrients and poisons, and to prey and predators. Given the importance of categorization, it is not surprising that many different cognitive theories of human category learning have been proposed and tested. Although some of these theories have unquestionably been more successful than others, most categorization researchers would probably acknowledge disappointment with how difficult it has been to differentiate the predictions of the more successful models, despite the very different cognitive processes they hypothesize. It appears that behavioral category-learning data do not offer enough constraints to identify the correct model.

The cognitive neuroscience revolution has offered some exciting new tools for resolving these conflicts. In particular, the past decade has seen an explosion of new results that collectively are beginning to paint a detailed picture of the neural mechanisms and pathways that mediate category learning. These results come from a wide variety of sources, including animal lesion, single-cell recording, neuroimaging, and neuropsychological studies. Lagging somewhat behind this avalanche of new data has been the development of new theories that can account for the traditional cognitive results as well as for these newer neuroscience results. Even so, some such theories have been proposed. The most comprehensive and best tested of these is the competition between verbal and implicit systems (COVIS) theory of category learning [Ashby et al. (1998); Ashby and Waldron (1999)]. This chapter describes COVIS and many of the cognitive experiments that have been run to test some of its strong and surprising \textit{a priori} predictions. The next chapter, entitled “Multiple Systems of Perceptual Category Learning: Neuropsychological Tests,” describes tests of COVIS using neuropsychological patient data and neuroimaging results.

Briefly, COVIS postulates two systems that compete throughout learning – a frontal-based explicit system that uses logical reasoning and depends on working memory and executive attention, and a basal ganglia-mediated implicit system that uses procedural learning. The procedural learning-based system is phylogenetically older. It can learn a wide variety of category structures, but it learns in a slow, incremental fashion and is highly dependent on reliable and immediate feedback. In contrast, the explicit, rule-based system can learn a fairly small set of category structures quickly – specifically, those structures that can be learned via an explicit reasoning process. Tasks that require subjects to learn such structures are called \textit{rule-based category-learning tasks}. On the other hand, there are many category structures that the explicit system cannot learn. An important example occurs in \textit{information-integration tasks}, in which learning requires subjects to integrate perceptual information across two or more noncommensurable stimulus dimensions. Before describing COVIS in detail, we take a short detour to describe rule-based and information-integration category-learning tasks.
2. Two Category-Learning Tasks

As mentioned above, rule-based tasks are those in which the categories can be learned via some explicit reasoning process. Frequently, the rule that maximizes accuracy (i.e., the optimal strategy) is easy to describe verbally. In the most common applications, only one stimulus dimension is relevant, and the subject’s task is to discover this relevant dimension and then to map the different dimensional values to the relevant categories. However, there is no requirement that rule-based tasks be one-dimensional. For example, a conjunction rule (e.g., respond A if the stimulus is small on dimension x and small on dimension y) is a rule-based task because a conjunction is easy to describe verbally.

Information-integration category-learning tasks are those in which accuracy is maximized only if information from two or more noncommensurable stimulus dimensions is integrated at some predecisional stage [Ashby and Gott (1988)]. Perceptual integration could take many forms – from computing a weighted linear combination of the dimensional values to treating the stimulus as a Gestalt. Typically, the optimal strategy in information-integration tasks is difficult or impossible to describe verbally. Real-world examples of information-integration tasks are common. For example, deciding whether an X-ray shows a tumor requires years of training and expert radiologists are only partially successful at describing their categorization strategies.

Some stimuli from rule-based and information-integration tasks that are typical of the experiments described later in this chapter are shown in Figure 1. In both tasks, the two contrasting categories consist of circular sine-wave gratings (i.e., disks in which luminance varies sinusoidally). The disks are all of equal diameter, but they differ in bar width and orientation. In most experiments, each category contains hundreds of these disks. On each trial of a typical experiment, a disk is randomly selected and presented to the subject, whose task is to assign it to category A or B. In most cases, feedback about response accuracy is given at the end of each trial, and this process is repeated for hundreds or thousands of trials. Figure 1 also shows the decision bounds that maximize categorization accuracy.

In the rule-based task (Figure 1A), the optimal bound, denoted by the vertical line, requires subjects to attend to bar width and ignore the grating’s orientation. This bound has a simple verbal description: “Respond A if the bars are thin and B if they are thick.” In the information-integration task (Figure 1B), equal attention must be allocated to both stimulus dimensions. In this task, there is no simple verbal description of the optimal decision bound. For a more thorough discussion of rule-based and information-integration tasks, see Ashby and Maddox (2005).

3. COVIS

As mentioned earlier, COVIS postulates two systems that compete throughout learning – an explicit system that uses logical reasoning and an implicit system that uses a form
Fig. 1. (A) Stimuli that might be used in a rule-based category-learning task. The vertical line is the optimal category boundary. (B) Stimuli that might be used in an information-integration category-learning task. The diagonal line is the optimal category boundary.
of procedural learning. The explicit system generates and tests hypotheses about category membership, a process that requires working memory and executive attention, and it controls performance in rule-based tasks, primarily because the optimal rule in these tasks is easy to reason about logically. The procedural learning that forms the basis of the implicit system has previously been associated with motor learning [e.g., Willingham, Nissen and Bullemer (1989), Willingham (1998)]. COVIS assumes that the implicit system learns in an incremental fashion and, unlike the explicit system, is not constrained to learning any particular class of categorization rules. In information-integration tasks, the optimal rule is difficult to describe verbally, so it is unlikely to be discovered by the hypothesis generation process of the explicit system. As a result, COVIS assumes that the procedural-learning system dominates (asymptotic) performance in information-integration tasks.

The next two major sections describe the COVIS explicit and implicit systems, respectively. Computational versions of the two systems are described in the appendix.

4. The COVIS explicit system

The COVIS explicit system assumes that subjects generate and test hypotheses about category membership. For example, the initial hypothesis may be to “respond A if the grating is tilted up, and B if it is tilted down.” This candidate rule is then held in working memory while it is being tested. With the categories shown in Figure 1A, performance based on this rule will be suboptimal, so the feedback will eventually tell the subject that this hypothesis is incorrect. At this point, an alternative hypothesis must be selected, and executive attention must be switched from the old hypothesis to the new. This selection, switching, and testing process continues until performance is satisfactory, or until the subject gives up and decides that no satisfactory rule exists.

Figure 2 shows the neural mechanisms that mediate performance in the COVIS explicit system during a trial of the category-learning task illustrated in Figure 1. The key structures in the model are the anterior cingulate, the prefrontal cortex (PFC), and the head of the caudate nucleus. There are two separate subnetworks in this model – one that maintains candidate rules in working memory during the testing process and that mediates the switch from one rule to another, and one that generates or selects new candidate hypotheses.

The working memory maintenance and attentional switching network includes all structures in Figure 2, except the anterior cingulate. The idea is that the long-term representation of each possible salient rule is encoded in some neural network in sensory association cortex. These cortical units send excitatory signals to working memory units in lateral PFC, which send recurrent excitatory signals back to the same cortical units, thereby forming a reverberating loop. At the same time, the PFC is part of a second excitatory reverberating loop through the medial dorsal nucleus of the thalamus [Alexander, Delong and Strick (1986)]. These double reverberating loops maintain activation in the PFC working memory units during the hypothesis-testing procedure.
One potential problem with this scheme is that an inhibitory input to the thalamus from the globus pallidus could disrupt processing in the reverberating loops. The spontaneous activity in the GABAergic pallidal cells is high, so without some intervention, this tonic inhibition will prevent the thalamus from closing the cortical-thalamic loop, and the information will be lost from working memory. The solution to this problem is for the PFC to excite the head of the caudate nucleus. Cells in the caudate have a low spontaneous firing rate, and instead are characterized by bursts of activity when stimulated by the cortex [Bennett and Wilson (2000)]. The direct excitatory projection from the PFC causes the caudate cells to fire bursts when the working memory cells in the lateral PFC are active. These bursts of caudate activity inhibit the pallidal cells (since the caudate cells are GABAergic), thereby preventing them from disrupting the reverberating cortical-thalamic working memory loop.

When feedback convinces a subject that he or she is using an incorrect rule, then a new rule must be selected and executive attention must be switched from the old rule to this new one. In COVIS, separate processes mediate these selection and switching operations. The process of generating new candidate hypotheses is clearly complex, and a complete model of rule selection is beyond the scope of this chapter, in part because it likely involves poorly understood phenomena such as creativity and insight. We largely beg the question of exactly how rule selection occurs – that is, we simply
assume that some cortical network that includes the PFC and the anterior cingulate mediates the process. In the computational version of the model, the anterior cingulate selects among alternative rules by enhancing the activity of the specific PFC working memory unit that represents a particular rule [Ashby, Valentin and Turken (2002b)].

Once a new rule is selected, its representation is maintained in its own set of reverberating loops. In Figure 2, one such loop maintains the representation of a rule focusing on the orientation of the bars and one loop maintains a rule focusing on the width. The next stage is to switch attention to the loop encoding this new rule. The following section describes this switching operation in some detail.

4.1. Switching attention in the explicit system

Switching, by definition, involves a change in the focus of executive attention. Thus, any model of switching must make assumptions about attention. There is growing consensus that the human attention system is subserved by separate subsystems, or at least by a broad anatomical network in which different subtasks are mediated by different brain areas [e.g., Posner and Petersen (1990), Olshausen, Anderson and Van Essen (1993), LaBerge (1997)]. Although a number of different theories have been proposed, there is widespread agreement that perceptual (i.e., visual) attention is mediated by a posterior system that includes visual cortex, much of the PPC, the pulvinar, and the superior colliculus [e.g., Posner and Petersen (1990), Olshausen et al. (1993)]. In contrast, executive (i.e., conscious) attention is thought to be mediated by an anterior system that includes the anterior cingulate, PFC, and perhaps the basal ganglia and pulvinar [e.g., Posner and Petersen (1990), LaBerge (1997)].

Although these systems are anatomically separate, the question of whether they are functionally separate has been only recently addressed. The actions of the two systems are often correlated, yet empirical evidence that perceptual and executive attention can function independently is growing [e.g., Pashler (1991), Johnston, McCann and Remington (1995), Maddox, Ashby and Waldron (2002)]. Johnston et al. (1995) provided some of the strongest evidence to date for the functional independence of perceptual and executive attention by showing that perceptual and executive attention systems operate at different temporal stages of processing. Specifically, they showed that a critical reference stage of processing (letter identification in their task) operates after the perceptual attention stage, but prior to the executive attention stage.

Although it is common to assume that visual attention can be allocated simultaneously to more than one location [e.g., Palmer, Vergese and Powel (2000)], the possible functional independence of perceptual and executive attention means that executive attention, which we assume to operate in the PFC, might follow different rules. In fact, we assume that only one rule (or item) can be the current focus of attention. A variety of evidence supports this assumption [e.g., Baars (1988)]. For example, a long series of studies shows that, with sequentially presented items, rapid privileged access is available only for the single most recent item [Dosher (1981), McElree (2001)]. Even so, it is important to acknowledge that this assumption is controversial. In particular, there is a current debate
as to whether executive attention can be allocated to a single item [e.g., Baars (1988), McElree and Dosher (2000)] or simultaneously to multiple items [e.g., Cowan (2000)]. For our purposes, this issue is not critical. We assume that only one rule can be within the current focus of attention, but an alternative version of the COVIS explicit system could easily be developed that allows executive attention to focus on several rules simultaneously. In either case, like a variety of models in the literature [e.g., Cowan (1988)], we assume a hierarchical organization of working memory. The most privileged set is the single rule in the current focus of attention. Next is a larger set of rules that can quickly become the focus of attention. These are candidate rules, each with their own reverberating working memory loops through the PFC. A much larger set of possible rules is represented in long-term memory and, if properly activated, could be moved into the intermediate set – that is, they could activate a PFC working memory unit and initiate their own reverberating circuit.

There are two general types of attentional switches – automatic and volitional. Automatic attentional switches are typically initiated by some unexpected salient event that occurs in the environment. In contrast, volitional switches are initiated internally via some intentional process that requires effort. In category-learning tasks, attentional switching is volitional, so this section focuses on volitional switches. This is an important provision because we believe that automatic and volitional switches are mediated differently in the mammalian brain. For example, the neuromodulator dopamine likely plays a much more critical role in automatic switches than in volitional switches. Evidence for this comes from a long line of studies showing that dopamine cells (e.g., in the VTA) fire in response to any unexpected salient stimulus, including unexpected rewards [Mirenowicz and Schultz (1994)], stressful or noxious stimuli [e.g., Imperato et al. (1991), Sorg and Kalivas (1993)], and even random light flashes and tones that are not associated with either reward or punishment [Horvitz, Stewart and Jacobs (1997)]. Any of these environmental events is likely to elicit an automatic attentional switch.

In COVIS, a volitional switch of attention from an old rule to a new rule is mediated by a reduction in the PFC excitatory input to the head of the caudate nucleus. Such deactivation would cause activation in the globus pallidus to increase back to its high baseline levels, which in turn would inhibit the thalamus, thereby breaking the cortical-thalamic loop. For example, the two classic symptoms of a PFC lesion are impaired working memory and perseveration on the Wisconsin Card Sorting Test [e.g., Kimberg, D’Esposito and Farah (1997)] – that is, a failure to switch volitionally when the feedback signals that a switch is necessary. Frontal patients have an intact dopamine system and intact basal ganglia. However, since they have lesions of the PFC, they presumably have decreased cortical control of the basal ganglia. So, the excitatory projection from the PFC to the head of the caudate represents a promising candidate via which a volitional switch might be mediated. In particular, with conscious effort, a person could initiate a switch by significantly decreasing this excitation, which would decrease the caudate output, increase the pallidal inhibition of the thalamus, and thereby disrupt the loop that is maintaining the working memory of the rule in the current focus of attention. On the other hand, with conscious
effort, a person could also allocate more attention to the current rule by increasing the PFC activation of the caudate, which in turn would increase the caudate output, decrease the pallidal output, and thereby strengthen the reverberation in the cortical-thalamic loop.

According to this hypothesis, the signal for a volitional switch originates in the PFC, but the switching itself is mediated within the basal ganglia (i.e., an increase in basal ganglia output breaks the cortical-thalamic loop). Evidence supporting the hypothesis that the basal ganglia play an important role in attentional switching comes from several sources. First, injections of a glutamate agonist directly into the striatum increase the frequency with which cats switch from one motor activity to another in a task where food rewards are delivered for such switching behaviors [Jaspers, De Vries and Cools (1990)]. Second, lesioning the dopamine fibers that project from the VTA into the PFC improves the performance of monkeys in an analogue of the Wisconsin Card Sorting Test [Roberts et al. (1994)]. If switching occurs in the PFC, then such lesions should impair switching performance. However, the PFC tonically inhibits the VTA (i.e., via a negative feedback loop). Lesioning the dopamine fibers projecting into the PFC releases the VTA from this inhibition. As a consequence, such lesions increase dopamine release into the basal ganglia [Roberts et al. (1994)]. If the basal ganglia are responsible for switching, and if switching is enhanced by dopamine, then lesioning dopamine fibers into the PFC should improve switching, which is exactly the result observed by Roberts et al. (1994). Third, Van Golf Racht-Delatour and El Massiou (1999) demonstrated that rats with lesions to the dorsal striatum had no deficits in learning which arm of a radial arm maze was initially baited, but they did have deficits, relative to rats with sham lesions, when the position of the baited arm was successively switched according to a simple rule. Finally, there are well-known switching deficits in individuals with caudate dysfunction. For example, numerous studies have shown that Parkinson’s disease patients, who have abnormally low levels of dopamine in the striatum, have a greater tendency to perseverate on the Wisconsin Card Sorting Test [Brown and Marsden (1988)].

4.2. Long-term storage of explicit category knowledge

Consider a case where the explicit system discovers the optimal categorization rule through this hypothesized process of hypothesis generation and testing. Of course, working memory cannot be used to maintain the representation of this optimal rule over any significant period of time. COVIS assumes that the consolidation and long-term representation of category knowledge obtained by the explicit system is mediated by medial temporal lobe structures, and in particular, by the hippocampus. Thus, as in many other models, the projection from the PFC to the hippocampus is assumed to mediate the transition from working memory to long-term declarative memory representations [e.g., Eichenbaum (1997)].

The medial temporal lobe’s declarative memory systems might play an even more prominent role in complex rule-based tasks. When there are many alternative explicit rules to consider, the process of sorting through all these possible rules may exceed the capacity
of working memory. In this case, episodic and semantic memory may be required to help
the subject keep track of which rules have already been tested and rejected. Note that this
hypothesis predicts that medial temporal lobe amnesiacs should learn normally in simple
rule-based tasks, but they should be impaired in complex rule-based tasks. Several studies
have reported evidence supporting the former prediction [Leng and Parkin (1988),
Janowsky et al. (1989)], but to our knowledge the latter prediction has not been tested.

5. The COVIS procedural-learning system

Figure 3 shows the COVIS procedural learning-based system [Ashby et al. (1998),
Ashby and Waldron (1999)]. The key structure is the caudate nucleus, a major input
region within the basal ganglia. In primates, all of the extrastriate visual cortex projects
directly to the tail of the caudate nucleus, with about 10,000 visual cortical cells con-
verging on each caudate cell [Wilson (1995)]. The model assumes that, through a pro-
cedural learning process, each caudate unit associates an abstract motor program with
a large group of visual cortical cells (i.e., all the ones that project to it).

The medium spiny cells in the tail of the caudate send projections to a variety of pre-
frontal and premotor cortical areas. There are two synapses on this pathway. The first

Fig. 3. The COVIS procedural-memory-based category-learning system. (Excitatory projections
end in solid circles, inhibitory projections end in open circles, and dopaminergic projection is dashed.
PFC = prefrontal cortex, Cau = caudate nucleus, GP = globus pallidus, and Th = thalamus.)
synapse on the principal path is in the globus pallidus, which is a major output structure within the basal ganglia. The second synapse is in the thalamus, primarily in the ventral anterior nucleus, pars magnocellularis (VAmc). The primary cortical projection from VAmc is to Brodmann Area 8 and the so-called supplementary eye fields [Shook, Schlag-Rey and Schlag (1991)].

Area 8 lies within the anterior bank of the arcuate sulcus, which falls between the PFC and the premotor cortex. Although some authors consider Area 8 to be transitional in function between the prefrontal and premotor cortex [e.g., von Bonin and Bailey (1947)], more recent classifications include it as a premotor region [e.g., Pandya and Yeterian (1985), Passingham (1993)]. The supplementary eye fields, however, are part of the supplementary motor area (SMA) (Brodmann Area 6). Area 8 and the supplementary eye fields both mediate the selection of eye movements and orienting responses [Passingham (1993)]. In contrast, the SMA is primarily responsible for selecting motor programs to move limbs. So, for example, animals with SMA lesions are impaired in learning to move a lever one way in response to one stimulus (call it A) and another way in response to a different stimulus (call it B), but animals with Area 8 lesions learn normally in this task. In contrast, animals with SMA lesions are unimpaired if the task involves learning to direct attention to a stimulus in one location if A is presented and to another stimulus in a different location if B is presented, whereas animals with Area 8 lesions are impaired [Halsband and Passingham (1985), Petrides (1985)]. These results suggest that the COVIS implicit system may learn orienting responses, rather than abstract category labels. Few studies have directly tested this prediction, but consistent eye movements have been observed in a variety of cognitive tasks that had no objective requirement for subjects to move their eyes [e.g., Spivey and Geng (2001)].

COVIS assumes that procedural learning in the caudate is facilitated by a dopamine-mediated reward signal from the substantia nigra (pars compacta). There is a large literature linking dopamine and reward, and many researchers have argued that a primary function of dopamine is to serve as the reward signal in reward-mediated learning [e.g., Miller, Sanghera and German (1981), Wickens (1993), Montague, Dayan and Sojnowski (1996)]. Fairly specific neurobiological models of this learning process have been developed [e.g., Wickens (1993), Calabresi et al. (1996)].

Figure 4 shows a close-up view of a synapse between the axon of a pyramidal cell originating in the visual cortex and the dendrite of a medium spiny cell in the caudate nucleus. Note that glutamate projections from the visual cortex and dopamine projections from the substantia nigra both synapse on the dendritic spines of caudate medium spiny cells [e.g., DiFiglia, Pasik and Pasik (1978), Freund, Powell and Smith (1984), Smiley et al. (1994)]. There are a number of different glutamate receptors, but for our purposes the two most important are the NMDA and AMPA receptors. The AMPA receptor becomes active when small amounts of glutamate are released presynaptically. However, the NMDA receptor has a high threshold for activation because, when the postsynaptic cell is hyperpolarized, the NMDA receptor is blocked by a magnesium plug. This plug dissociates from the receptor after the cell is partially depolarized. Thus, a strong presynaptic glutamate signal is required to activate postsynaptic NMDA receptors.
The best available evidence indicates that three factors are required to strengthen synapses of the type shown in Figure 4 (i.e., for long-term potentiation (LTP) to occur): (1) strong presynaptic activation, (2) strong postsynaptic (i.e., NMDA receptor) activation, and (3) dopamine release [e.g., Nairn et al. (1988), Wickens (1993), Calabresi et al. (1996)]. The first factor should occur for those visual cortical cells that are maximally tuned to the presented stimulus, and the third factor will occur if the subject is rewarded for a correct response. Thus, according to COVIS, synapses in the implicit system between visual cortical cells and medium spiny cells in the caudate are strengthened if the visual cortical cell is maximally tuned to the presented stimulus (factors 1 and 2) and the subject is rewarded for responding correctly (factor 3).

On the other hand, if either of factors 2 or 3 is missing, then it is thought that the strength of the synapse will weaken [i.e., long-term depression (LTD) will occur; e.g., Calabresi et al. (1996)]. Note that there are several ways this could happen. One is if the subject responds incorrectly (so factor 3 is missing), and another is if the visual cortical cell responds only weakly to the presented stimulus. Thus, this model of LTD predicts that any synapse responsible for the subject emitting an incorrect response will be weakened, as will any synapse that is driven by a cell in the visual cortex that does not encode the perceptual representation of the stimulus. The combination of this with the three-factor model of LTP produces a powerful learning algorithm.

The three-factor model of LTP is appealing, but a serious timing problem must be solved before it can operate effectively. The problem is that shortly after the stimulus is presented, the (visual) cortical-striatal (i.e., tail of caudate) synapse will be activated, but the
dopamine release must necessarily occur some time later (e.g., several seconds), because it follows the reward, which follows the response, which follows the stimulus presentation. Evolution has produced a beautiful and ingenious solution to this problem. NMDA receptor activation causes an influx of free Ca\(^{2+}\) into the spines of the caudate medium spiny cells. This Ca\(^{2+}\) triggers a number of chemical reactions, some of which have the effect of depolarizing the cell and eventually causing it to fire. After the cell fires, a natural hyperpolarization process is triggered that resets its membrane potential. The result is that by the time the dopamine is released, the depolarization produced by the presynaptic glutamate release has been erased from the major compartments of the medium spiny cell. However, because the spines are somewhat separated from the bulk of the intracellular medium, the mechanisms that reset the membrane potential operate more slowly than in the main cellular compartments. In fact, it turns out that free Ca\(^{2+}\) persists in the spines for several seconds after entering the cell [MacDermott et al. (1986), Gamble and Koch (1987)]. Thus, so long as the reward is delivered within a few seconds of the response, a trace will still exist in the critical spines that were responsible for eliciting the behavior that earned the reward, and so the correct synapses will be strengthened (i.e., via LTP). Note that an obvious, and exceptionally strong prediction of this model is that if the feedback is delayed more than a few seconds, then learning should be severely disrupted in information-integration tasks.

It is important to note that the model shown in Figure 3 is strictly a model of visual category learning. However, it is feasible that a similar system exists in the other modalities, since almost all of them also project directly to the basal ganglia, and then indirectly to frontal cortical areas [via the thalamus and either the substantia nigra pars reticulata or the globus pallidus; e.g., Chudler, Sugiyana and Dong (1995)]. The main difference is in where within the basal ganglia they initially project. For example, auditory cortex projects directly to the body of the caudate [i.e., rather than to the tail; Arnalud et al. (1996)].

6. Competition between the COVIS explicit and implicit systems

If human category learning is mediated by multiple systems, then an important question is how the various systems interact and how their separate contributions are coordinated during the response selection process. There are almost no data that directly address this question, and as a result this topic is likely to be the focus of intense research efforts during the next few years. In one of the few studies examining this issue, Poldrack et al. (2001) reported an antagonistic relationship between neural activation in the basal ganglia and medial temporal lobes (i.e., using fMRI) – that is, basal ganglia activation tended to increase with category learning, whereas medial temporal lobe activation decreased. While this result is suggestive, it is obviously difficult to draw strong conclusions from a single study.

The original version of COVIS [Ashby et al. (1998)] assumed that the explicit and implicit systems learned independently, but that they competed for control of the observable response on each trial. The competition was implemented in the following way. The output of each COVIS system is a discriminant value. Roughly speaking,
these discriminant values measure the distance from the percept to the decision bounds used by each system, with the provision that the values are positive in the response region assigned to category A (say) and negative in the region assigned to category B. Response confidence generally increases with distance to bound, so the absolute value of the discriminant is a measure of confidence. In COVIS, parameters $\theta_E$ and $\theta_I$ measure the relative strengths of the explicit and implicit systems, respectively (where $\theta_E + \theta_I = 1$). COVIS assumes that people initially rely almost exclusively on the explicit system (at least in western cultures), so at the beginning of learning $\theta_E$ is much larger than 0.5 (our applications typically set the initial value of $\theta_E$ to 0.99). However, in information-integration tasks, the explicit system frequently fails and the implicit system is often rewarded, and as a consequence $\theta_E$ gradually decreases with practice and $\theta_I$ increases. On each trial, the observable response is determined by whichever system has the greatest weighted confidence, which is defined as the system weight ($\theta_E$ or $\theta_I$) multiplied by the absolute value of the system discriminant function.

Simulations of COVIS in information-integration tasks show that after extensive practice $\theta_I$ exceeds $\theta_E$, but even after thousands of trials, $\theta_E$ rarely falls much below 0.4. Thus, COVIS predicts that even after massive amounts of practice, subjects will still use explicit rules on some trials in information-integration tasks. For example, in the Figure 1B information-integration task, disks with the narrowest bars are all in category A and disks with the widest bars are all in category B. COVIS predicts that for such extreme stimuli subjects may apply an explicit rule, whereas for disks with bars of intermediate width they would learn that no rule is effective and instead rely on procedural knowledge.

7. Dissociations between rule-based and information-integration category learning

A number of recent experiments have tested parameter-free a priori predictions made by COVIS. Collectively, the results of these studies also provide strong evidence that learning in these tasks is mediated by separate systems. A number of these results test COVIS predictions about the nature and timing of trial-by-trial feedback concerning response accuracy. In particular, COVIS predicts that, because the explicit system has access to working memory and executive attention, it should be relatively unaffected by changes in the timing and form of the feedback signal. In contrast, the COVIS implicit system is highly sensitive to feedback parameters. First, because COVIS predicts that a dopamine-mediated reward signal is necessary for learning (e.g., LTP) to occur in the caudate nucleus, the absence of such a reward signal should greatly interfere with this form of implicit category learning. In support of this prediction, in the absence of any trial-by-trial feedback about response accuracy, people can learn some rule-based categories, but there is no evidence that they can learn information-integration categories [Ashby, Queller and Berretty (1999)].

Second, COVIS predicts that learning in information-integration tasks should be impaired (relative, say, to learning in rule-based tasks) when the category label is shown before stimulus presentation (rather than after the response). To test this prediction,
Ashby, Maddox and Bohil (2002a) trained subjects on rule-based and information-integration categories using an observational training paradigm in which subjects are told before stimulus presentation which category the ensuing stimulus will be from. Following stimulus presentation, subjects then pressed the appropriate response key. Traditional feedback training was as effective as observational training with rule-based categories, but with information-integration categories, feedback training was significantly more effective than observational training.

A third feedback-related prediction of COVIS is that the timing of the feedback signal relative to the response should be critical for information-integration learning but not for rule-based learning. As described above, if the reward (e.g., feedback signal) is delayed by more than a few seconds, then the ensuing dopamine release will strengthen inappropriate synapses in the tail of the caudate and learning in information-integration tasks should therefore be adversely affected. In contrast, the explicit system can hold critical information in working memory for many seconds, and therefore COVIS predicts that rule-based learning should not be affected significantly by feedback delays of a few seconds. These predictions were tested by Maddox, Ashby and Bohil (2003), who showed that information-integration category learning is impaired if the feedback signal is delayed by as little as 2.5 s after the response. In contrast, delays as long as 10 s had no effect on rule-based category learning.

Another set of studies tested the fundamental assumption of COVIS that information-integration categorization uses procedural learning, whereas rule-based category learning does not. Ashby, Ell and Waldron (2003) had subjects learn either rule-based or information-integration categories using traditional feedback training. Next, some subjects continued as before, some switched their hands on the response keys, and for some the location of the response keys was switched (so the Category A key was assigned to Category B and vice versa). For subjects learning rule-based categories, there was no difference among any of these transfer instructions, suggesting that abstract category labels are learned in rule-based categorization. In contrast, for subjects learning information-integration categories, switching hands on the response keys caused no interference, but switching the locations of the response keys caused a significant decrease in accuracy. Thus, it appears that response locations are learned in information-integration categorization, but specific motor programs are not. The importance of response locations in information-integration category learning, but not in rule-based category learning, was confirmed in a recent study by Maddox, Bohil and Ing (2004b). These information-integration results essentially replicate results found with traditional procedural-learning tasks [Willingham et al. (2000)].

A third set of studies tested the COVIS prediction that working memory and executive attention are critical for rule-based category learning, but not for information-integration category learning. First, Waldron and Ashby (2001) had subjects learn rule-based and information-integration categories under typical single-task conditions and while simultaneously performing a secondary task that requires working memory and executive attention. The dual task had a massive detrimental effect on subjects’ ability to learn the simple one-dimensional rule-based categories (trials to criterion increased by 350%), but no significant effect on their ability to learn the complex information-integration categories. This result alone is highly problematic for unified
accounts of rule-based and information-integration categorization. Arguably, the most successful existing single-process model of category learning is Kruschke’s (1992) exemplar-based ALCOVE model. Ashby and Ell (2002) showed that the only versions of ALCOVE that can fit the Waldron and Ashby data make the strong prediction that, after reaching criterion accuracy on the one-dimensional rule-based structures, subjects would have no idea that only one dimension was relevant in the dual-task conditions. Ashby and Ell (2002) reported empirical evidence that strongly disconfirmed this prediction. Thus, the best available single-system model fails to account even for the one dissociation reported by Waldron and Ashby (2001).

Second, Maddox et al. (2004a) tested the COVIS prediction that feedback processing requires attention and effort in rule-based category learning, but not in information-integration category learning. In this study, subjects alternated a trial of categorization with a trial of Sternberg (1966) memory scanning. The two conditions were identical except that one had a short delay after the categorization trial and before memory scanning, whereas the other had a short delay after memory scanning and before categorization. Information-integration category learning was identical in these two conditions, whereas rule-based category learning was significantly impaired when subjects had only a short time to process the categorization feedback.

It is important to realize that these dissociations are not driven simply by differences in the difficulty of rule-based versus information-integration tasks. First, in several cases the experimental manipulation interfered more with the learning of the simple rule-based categories than with the more difficult information-integration strategies [Waldron and Ashby (2001), Maddox et al. (2004a)]. Second, most of the studies explicitly controlled for difficulty differences either by decreasing the separation between the one-dimensional rule-based categories, or by using a more complex two-dimensional conjunction rule in the rule-based conditions. Both manipulations increase the difficulty of rule-based categorization, yet in no case did such increases in difficulty affect the qualitative dissociations described above.

8. Conclusions

The preceding section demonstrates that COVIS provides a powerful description of the differences between the learning that dominates in rule-based and information-integration category-learning tasks. Even more impressive, however, is the fact that without COVIS, many of the experiments described above would never have been run. For example, it is difficult to imagine how any purely cognitive theory could ever predict that delaying feedback by a few seconds should interfere with one type of category learning but not with another, or that switching the location of the response keys should cause a similar selective interference. Of course, one of the greatest benefits of a good theory is that it suggests new experiments to run, which themselves add to our knowledge in new and unexpected ways. In this sense, at least, COVIS has already been quite successful.

This chapter has focused on cognitive behavioral tests and COVIS’s predictions. But the theory also makes a variety of other types of predictions – mostly because of
its neurobiological underpinnings. In particular, it makes predictions about which neuropsychological patient groups should be impaired in category learning, and which groups should have normal category-learning abilities. Further, when it predicts a deficit, it also makes fairly specific predictions about its precise nature. For example, COVIS predicts that patients with lesions confined to the PFC should be impaired in rule-based category-learning tasks, but relatively normal in information-integration tasks. The next chapter reviews the existing neuropsychological tests of COVIS, with a focus on patients with striatal or hippocampal dysfunction.

Despite its many initial successes, there are some recent signs that COVIS may be incomplete as a theory of human category learning. One important body of evidence pointing in this direction comes from a third type of category-learning task called the “(A, not A) prototype distortion task” [Ashby and Maddox (2005)], in which each exemplar of one category is created by randomly distorting a single prototype [e.g., Posner and Keele (1968)]. The subject’s task is to respond “Yes” or “No” depending on whether or not the presented stimulus is a member of this category (not-A stimuli are just random patterns). Several studies have shown that a variety of neuropsychological patient groups that are known to have deficits in rule-based and/or information-integration tasks show apparently normal (A, not A) prototype distortion learning, a result that is obviously problematic for the original version of COVIS. This list includes patients with Parkinson’s disease [Reber and Squire (1999)], schizophrenia [Kéri et al. (2001)], or Alzheimer’s disease [Sinha (1999), also see Kéri et al. (1999)]. COVIS hypothesizes that rule-based category learning uses working memory and other declarative memory systems (e.g., episodic and semantic memory), and that information-integration learning uses procedural memory. One possibility is that performance in (A, not A) prototype distortion tasks is mediated by some different memory system. An obvious possibility is the perceptual representation memory system [Reber and Squire (1999), Ashby and Casale (2002)]. In support of this hypothesis, all of the existing neuroimaging studies of (A, not A) prototype distortion tasks have reported learning-related changes in the occipital cortex [Reber, Stark and Squire (1998a,b), Aizenstein et al. (2000)]. Learning is, by definition, the process of laying down some sort of memory trace, and there is certainly no reason to suspect that any of the separate memory systems that have been hypothesized are incapable of storing memories about categories. For this reason, a complete theory of human category learning is likely to assign some role to each of the different memory systems that have been identified by memory researchers.

Appendix A

A.1. Network implementation of the explicit system

This section describes the computational version of the COVIS explicit system. The model described in this section is a hybrid neural network that consists of both symbolic and connectionist components. The model’s hybrid character arises from its
combination of explicit rule selection and switching and its incremental salience-learning component.

First, denote the set of all available explicit rules by $R = \{R_1, R_2, \ldots, R_m\}$. In most applications, the set $Y$ will include all possible one-dimensional rules, and perhaps a variety of plausible conjunction rules. Next, suppose rule $R_i$ is used on trial $n$. If the response on trial $n$ is correct, then rule $R_i$ is used again on trial $n + 1$. If the response on trial $n$ is incorrect, then the active rule on trial $n + 1$ is selected via the following three steps.

**Step 1**: Choose a rule at random from $Y$. Call this rule $R_j$. This is the rule selected by the frontal selection system (i.e., anterior cingulate/PFC) for a possible attentional switch.

**Step 2**: Define a weight, $Y$, for each rule as follows: $Y_i(n) = Z_i(n) + \gamma$, for the active rule $R_i$, $Y_j(n) = Z_j(n) + X$, for the rule $R_j$ chosen in step 1, and $Y_k(n) = Z_k(n)$, for all rules $R_k \neq R_i$ or $R_j$.

The constant $Z_i(n)$ is a measure of the current salience of rule $R_i$. The parameter $\gamma$ is a positive constant that reflects the subject’s tendency to perseverate. The larger the value of $\gamma$, the greater the tendency to perseverate on the current rule, even in the presence of negative feedback. Thus, $\gamma$ reflects the ability of the head of the caudate to switch attention to a new rule. If $\gamma$ is small, then switching will be easy, whereas switching is difficult if $\gamma$ is large. Finally, $X$ is a positive-valued random variable that represents the frontal selection system’s attempt to select rule $R_j$. In most applications, $X$ is a Poisson-distributed random variable with mean $\lambda$. Larger values of $\lambda$ increase the probability that attention will be switched to the selected rule $R_j$.

The initial salience of rule $R_i$ [i.e., $Z_i(0)$] is set to a value that reflects the subject’s initial tendency to experiment with this rule. For example, one reasonable strategy would be to set the initial salience of all one-dimensional rules to one constant value and the initial salience of any conjunction rules to some lower constant value. After setting these initial saliences, they are then updated each time rule $R_i$ is used. Specifically, if rule $R_i$ is used on trial $n$–1 and a correct response occurs, then

$$Z_i(n) = Z_i(n - 1) + \Delta_C,$$

where $\Delta_C$ is some positive constant. If rule $R_i$ is used on trial $n$ – 1 and an error occurs, then

$$Z_i(n) = Z_i(n - 1) - \Delta_E,$$

where $\Delta_E$ is also a positive constant.

**Step 3**: Choose the rule for trial $n + 1$ with the greatest weight $Y$ – that is, choose rule $R_i$ on trial $n + 1$ if $Y_i(n) = \max[Y_j(n), Y_2(n), \ldots, Y_m(n)]$.

This algorithm has a number of attractive properties. First, the more salient the rule, the higher the probability that it will be selected, even after an incorrect trial. This feature enables the algorithm to be resistant to catastrophic interference. Second, after the first trial, feedback is used to adjust the selection probabilities up or down, depending on the success
of the rule type. Third, the model has separate selection and switching parameters, reflecting COVIS’s assumption that these are separate operations. The random variable $X$ models the selection operation. The greater the mean of $X$ (i.e., $\lambda$), the greater the probability that the selected rule ($R_j$) will become active. COVIS assumes $\lambda$ increases with dopamine levels in frontal cortex. In contrast, the parameter $\gamma$ models switching, because when $\gamma$ is large, it is unlikely that the system will switch to the selected rule $R_j$. COVIS assumes that switching will be more difficult when there are abnormally low levels of dopamine in the striatum. Accordingly, $\gamma$ is assumed to vary inversely with dopamine levels in the caudate.

Once a particular rule is selected, say, rule $R_j$, then the explicit system selects a response by computing a discriminant value $h_j(n)$ and using the following decision rule:

Respond A on trial $n$ if $h_j(n) > 0$; respond B if $h_j(n) < 0$.

A full description of how discriminant values are computed is beyond the scope of this chapter [for details, see Ashby (1992)], but the general idea, which is approximately correct, is that $h_j(n)$ is the distance from the stimulus to the decision bound, with the provision that these distances are positive in the response region associated with category A and negative in the region associated with category B.

Each one-dimensional rule has an associated decision criterion (e.g., the $x$-intercept of the vertical bound in Figure 1A), and each conjunction rule has two. The values of these constants are learned via a conventional gradient-descent process with learning rate $\delta$. Thus, the model has five parameters, $\gamma$ (perseveration parameter), $\lambda$ (selection parameter), $\Delta_C$ (salience increment when correct), $\Delta_E$ (salience increment when incorrect), and $\delta$ (gradient-descent learning rate).

A.2. Network implementation of the implicit system

From a computational perspective, the procedural learning in the caudate could occur in one of two ways. One possibility is that the caudate learns a decision bound. In this case, learning could be modeled as a process via which parameters that describe the bound are updated as the subject gains experience with the categories. A second possibility is that the caudate learns to assign responses to regions of perceptual space. Under this scenario, the decision bound has no psychological meaning, and instead is simply the partition between regions associated with contrasting responses. Ashby and Waldron (1999) tested these two possibilities and found strong evidence against the former. In particular, their data disconfirmed the hypothesis that people learn decision bounds in information-integration tasks. Instead, the data are consistent with the hypothesis that the caudate learns to assign responses to regions of perceptual space. Ashby and Waldron (1999) proposed a computational model called the striatal pattern classifier (SPC) that implements this idea. The SPC, which is illustrated in Figure A.1, is a model of the COVIS procedural-learning system.

Figure A.1A shows the category structure of an information-integration task. In this case, however, the optimal decision bound is quadratic. A large literature shows that healthy young adults often eventually learn to respond in a nearly optimal fashion in experiments of this type [e.g., Ashby and Gott (1988), Ashby and Maddox (1990, 1992), Maddox and Ashby (1993)].
A simplified version of the SPC is illustrated in Figure A.1B for the information-integration category-learning experiment shown in Figure A.1A. The two-dimensional bar width/orientation space shown in Figure A.1B depicts the perceptual representation of the disks used in the Figure A.1A experiment. Thus, small regions in this space are associated with a distinct cell in some extrastriate visual area (i.e., the cell maximally stimulated when a particular disk is shown). The four large dots represent four different...
striatal units. Each striatal unit is associated with one of the two category responses, which creates four distinct regions in perceptual space. In Figure A.1B, two of those regions are associated with category A and two are associated with category B. The model assumes that each percept is mapped to the nearest striatal unit, and therefore that the subject emits the response associated with the striatal unit that is nearest the percept. Perceptual noise causes the percept elicited by each stimulus to vary randomly from trial to trial. All existing applications of the SPC have assumed that this perceptual noise has a multivariate normal distribution with mean 0 on each dimension and variance-covariance matrix equal to $\sigma^2 I$, where I is the identity matrix. The predicted probability that a subject will assign a particular stimulus to category A, say, is therefore equal to the volume, or integral under the distribution representing the perceptual effects of that stimulus throughout the category A response region. For a detailed description of an efficient algorithm for computing these integrals, see Ennis and Ashby (2003). On trials when a stimulus with coordinates $\mathbf{x}$ is presented, a discriminant value can be defined for the SPC by computing $\log P(R_B|\mathbf{x}) - \log P(R_A|\mathbf{x})$.

A fundamental feature of the SPC is the convergence it assumes between high-resolution perceptual space (i.e., in the visual cortex) and low-resolution decision space (i.e., in the tail of the caudate). One implication of this convergence assumption is that the SPC predicts that there is a limit on the complexity of the decision bound that can be learned in information-integration tasks. Although many studies have shown that people can learn linear or quadratic information-integration bounds, only a few studies have tested people's ability to learn bounds more complex than the quadratic bound shown in Figure A.1A. In support of the SPC prediction, all of these studies have reported that subjects fail to learn these complex bounds, despite many hundreds of practice trials [McKinley and Nosofsky (1995), Ashby, Waldron and Berkman (2001)].

Acknowledgment

Preparation of this chapter was supported by Public Health Service Grant MH3760. We thank Todd Maddox for his helpful comments.

References


THE NEUROPSYCHOLOGY OF PERCEPTUAL CATEGORY LEARNING*

W. TODD MADDOX

University of Texas

J. VINCENT FILOTEO

University of California, San Diego & Veterans Administration San Diego Healthcare System

Contents

Abstract 574
1. Introduction 575
2. Competition between verbal and implicit systems (COVIS) 575
3. Testing a priori Predictions of COVIS 578
4. Perceptual category learning in neurological patients 581
   4.1. Nonlinear information-integration category learning in amnesia 582
   4.1.1. Model-based analyses 584
   4.2. Nonlinear information-integration category learning in striatal-damaged patients 585
   4.3. Rule-based category learning in PD 587
   4.4. Further study of information-integration category learning in PD 591
   4.4.1. Brief summary of the results 593
5. General discussion 595
References 597

* This research was supported in part by National Institute of Health Grant R01 MH59196 to WTM, National Institute of Neurological Disorders and Stroke Grant R01 41372 to JVF, and a James McDonnell Foundation Grant.
Abstract

There is widespread agreement that multiple qualitatively different category learning systems mediate the learning of different category structures. Two systems that have received support are (1) a frontal-based explicit system that uses logical reasoning, depends on working memory and executive attention, and is mediated primarily by the anterior cingulate, the prefrontal cortex, and the head of the caudate; and (2) a basal ganglia-mediated implicit system that uses procedural learning and requires a dopamine reward signal. This chapter reviews a large body of work conducted in our laboratories that examines the details of the two proposed systems using neurological patients as experimental participants. Collectively, the studies suggest that the medial temporal lobes are little involved in category learning with large categories. They also suggest that, in striatal-damaged patients, the need to ignore irrelevant information is predictive of a rule-based category learning deficit, whereas the complexity of the rule is predictive of an information-integration category learning deficit.
1. Introduction

Category learning involves laying down a memory trace that can be used to improve the efficiency (i.e., accuracy and speed) of responding. It is now widely accepted that mammals have multiple memory systems [Schacter (1987), Squire (1992)], and this fact alone makes it reasonable to postulate that multiple category learning systems might also exist. This chapter reviews a body of work that suggests that perceptual category learning is characterized by multiple systems, each of which involves a set of diverse neurocognitive processes. This chapter builds upon the work outlined in the previous chapter by Ashby and Valentin, who reviewed a number of studies that tested a priori predictions from a recently proposed neurobiologically plausible multiple systems theory called the Competition between Verbal and Implicit Systems theory [COVIS; Ashby et al. (1998), Ashby and Waldron (1999), for a review, see Maddox and Ashby (2004)]. In each study, some experimental manipulation was introduced that was predicted a priori to affect processing in one system, but not the other. All of these studies used healthy young adults as participants. These studies provide a nice foundation, but the next step is to examine the systems in greater detail. One way to achieve this goal is to examine category learning in neurological patients with damage to specific brain areas. The aim of this chapter is to review a body of work conducted in our laboratories that examined category learning in various patient populations.

This chapter is organized as follows. First, we briefly introduce COVIS and the proposed underlying neurobiology. A more detailed description is offered in the previous chapter and in Maddox and Ashby (2004). The second section briefly reviews the qualitative dissociations introduced in the previous chapter. The third section reviews a body of work that examines category learning in various patient populations. The third section is subdivided into sections devoted to rule-based and information-integration category learning in patients with amnesia, Parkinson’s disease (PD), or Huntington’s disease (HD). The final section offers a brief summary and conclusions. It is important to note that this is not a substantive review of the field. Two excellent reviews are provided by Kéri (2003) and Poldrack and Packard (2003). Rather, this chapter reviews and integrates a large body of patient work conducted in our laboratories and others that takes a systematic empirical approach, supplemented by the application of a series of quantitative models, to the study of perceptual category learning.

2. Competition between verbal and implicit systems (COVIS)

A growing body of research suggests that the learning of different types of category structures is mediated by different systems with distinct but partially overlapping neurobiological substrates [Reber and Squire (1994), Pickering (1997), Erickson Kruschke (1998), Smith, Patalano and Jonides (1998), however, see Nosofsky and Johansen (2000), Ashby and Ell (2001, 2002), Maddox and Ashby (2004)]. One of the most successful multiple systems models of category learning, and the only one that specifies
the underlying neurobiology, is COVIS. COVIS postulates two systems that compete throughout learning – an explicit hypothesis-testing system that uses logical reasoning and depends on working memory and executive attention, and a procedural-learning-based system that relies more on incremental and feedback-learning processes. One intriguing aspect of the procedural-learning-based system is its association with the processes involved in motor performance [e.g., Willingham (1998), Hazeltine and Ivry (2002)], which leads to the important prediction that categories learned via a procedural-learning-based system should be closely linked to the motor response.

Much of the evidence for multiple category learning systems comes from two different types of categorization tasks. Rule-based category learning tasks are those in which the category structures can be learned via some explicit reasoning process that treats each stimulus dimension separately. Frequently, the rule that maximizes accuracy (i.e., the optimal rule) is easy to describe verbally [Ashby et al. (1998)]. For example, in Figure 1a, the stimuli (with one presented on each trial) are composed of a single line that varies in length and orientation across trials. Each symbol in Figure 1 denotes the length and orientation of one stimulus. Also shown in Figure 1 are the decision bounds that maximize categorization accuracy. In the rule-based task, the optimal bound requires observers to attend to length and ignore orientation. The vertical bound in Figure 1a corresponds to the rule: “Respond A if the line is short and B if it is long.”

Information-integration category learning tasks are those in which accuracy is maximized only if information from two or more stimulus components (or dimensions) is integrated at some predecisional stage [Ashby and Gott (1988)]. Perceptual integration could take many forms – from treating the stimulus as a Gestalt to computing a weighted linear combination of the dimensional values1. In many cases, the optimal rule in information-integration tasks is difficult or impossible to describe verbally [Ashby et al. (1998)]. The information-integration task in Figure 1b was generated by rotating the rule-based categories by 45°. Note that the information-integration rule is linear. Figure 1c depicts a case in which the information-integration rule is nonlinear. Category structures like these were used in the studies reviewed in the previous chapter and in some of the studies reviewed below.

COVIS assumes that learning in rule-based tasks is dominated by an explicit system that uses working memory and executive attention to generate and test hypotheses and is mediated primarily by the anterior cingulate, the prefrontal cortex, and the head of the caudate nucleus (see Figure 2 from the previous chapter). Learning in information-integration tasks is dominated by an implicit procedural-learning-based system, which is mediated largely within the tail of the caudate nucleus [see Figure 3 in Chapter 25; Ashby et al. (1998), Willingham (1998), Ashby and Ell (2001)]. (see the previous chapter for details).

---

1 A conjunction rule (e.g., respond A if the stimulus is small on dimension x and small on dimension y) is a rule-based task rather than an information-integration task because separate decisions are first made about each dimension (e.g., small or large) and then the outcome of these decisions is combined (integration is postdecisional, not predecisional).
Fig. 1. Stimuli and optimal decision bound from (a) rule-based, (b) linear information-integration, and (c) nonlinear information-integration categorization condition. Open squares denote category A items, and filled circles denote category B items.
3. Testing \textit{a priori} Predictions of COVIS

In this section we briefly review a number of studies that provided empirical tests of \textit{a priori} predictions derived from the proposed neurobiological underpinnings of COVIS. Because the hypothesis-testing system is under conscious control and has full access to working memory\textsuperscript{2} and executive attention, the placement and timing of the feedback signal should not be critical for rule-based category learning because this information can be held consciously in working memory. In contrast, a procedural-learning system that is mediated within the tail of the caudate nucleus would not be accessible to conscious awareness and is far removed from working memory. As a result, it would depend more heavily on the placement and timing of the feedback. As a test of these predictions, rule-based and information-integration category learning was compared across an observational training condition (in which observers were informed before stimulus presentation what category the ensuing stimulus would be from) and a traditional feedback training condition (in which the category label followed the response) [Ashby et al. (2002)], and across an immediate feedback condition (in which corrective feedback was provided immediately following the response) and a delayed feedback condition (in which corrective feedback was delayed by 2.5, 5, or 10 s following the response) [Maddox, Ashby and Bohil (2003)]. In line with the COVIS predictions, observational training and delayed feedback negatively impacted information-integration category learning but had little effect on rule-based category learning.

\textsuperscript{2} Many argue for the existence of an implicit form of working memory that may not be available to conscious awareness. When we use the term “working memory,” we refer to a conscious, verbalizable process.

![Fig. 2. Nonlinear information integration category structures and representative stimuli used in Filoteo et al. (2001a). Open squares denote category A stimuli and filled circles denote category B stimuli. The broken quadratic curve denotes the optimal decision bound.](image-url)
The implicit system in COVIS is assumed to be procedural-learning-based. The quintessential paradigm for studying procedural learning is the serial reaction time (SRT) task. In a typical SRT task, one of \( n \) stimuli is presented on each trial and each stimulus is associated with its own response key. The observer’s task is to press the relevant key as quickly as possible. A large response time improvement is observed when the stimulus sequence is repeated, even though the observer is unaware that a sequence exists.

The implicit system in COVIS is assumed to be procedural-learning-based. The quintessential paradigm for studying procedural learning is the serial reaction time (SRT) task. In a typical SRT task, one of \( n \) stimuli is presented on each trial and each stimulus is associated with its own response key. The observer’s task is to press the relevant key as quickly as possible. A large response time improvement is observed when the stimulus sequence is repeated, even though the observer is unaware that a sequence exists.

The implicit system in COVIS is assumed to be procedural-learning-based. The quintessential paradigm for studying procedural learning is the serial reaction time (SRT) task. In a typical SRT task, one of \( n \) stimuli is presented on each trial and each stimulus is associated with its own response key. The observer’s task is to press the relevant key as quickly as possible. A large response time improvement is observed when the stimulus sequence is repeated, even though the observer is unaware that a sequence exists.
Willingham et al. (2000) showed that changing the location of the response keys interferes with SRT learning even when the sequence of stimulus positions is unchanged. In addition, they showed that SRT learning is unaffected by changing the sequence of finger movements as long as the location of the response keys remains fixed. If the implicit system in COVIS is procedural-learning-based, then changing the location of the response keys should adversely affect learning in this system, and thus information-integration category learning, whereas changing the finger press associated with each category response should not. On the other hand, hypothesis-testing systems are not typically linked to a specific motor response, and should not be especially sensitive to procedures that change the mapping between category label and response location. Two studies directly tested these hypotheses. Ashby, Ell and Waldron (2003) examined rule-based and information-integration category learning using a training-transfer procedure. There were three conditions: control, hand-switch, and button-switch. In the control condition, the response key assigned to category A was pressed with the left index finger and the response key assigned to category B was pressed with the right index finger during both training and transfer. In the hand-switch condition, the hands were crossed during training so that the response key assigned to category A was pressed with the right index finger and the response key assigned to category B was pressed with the left index finger. During transfer, the hands were uncrossed on the response keys. In the button-switch condition, training was identical to that in the control condition, but during transfer the locations of the buttons were switched. For the rule-based task, hand switching and button switching had no effect on performance. For the information-integration task, on the other hand, button switching led to a decrement in performance, but hand switching did not. These results suggest that the hypothesis-testing system learns abstract category labels, whereas the procedural-learning system learns response positions.

In a related study, Maddox, Bohil and Ing (2005) examined rule-based and information-integration category learning across two conditions. In the fixed response location condition, the response key assigned to category A was pressed with the left index finger, and the response key assigned to category B was pressed with the right index finger. In the variable response location condition, the response key assigned to category A was pressed with the left index finger on half the trials, and with the right index finger on half the trials. In line with the predictions of COVIS, information-integration category learning, but not rule-based category learning, was adversely affected in the variable response location condition.

Experimental manipulations that adversely affect rule-based, but not information-integration learning, have also been observed. Waldron and Ashby (2001) showed that rule-based category-learning was disrupted more than information-integration category learning by the simultaneous performance of a task that required working memory and executive attention (a numerical Stroop task). In addition, Maddox et al. (2004a) showed that rule-based category learning was disrupted by a sequential memory-scanning task whereas information-integration category learning was not.

Taken together, these studies provide strong support for the existence of hypothesis-testing and procedural-learning-based systems of category learning and for the
neurobiological underpinnings proposed in COVIS. They provide an excellent first step
and help lay the groundwork for more detailed examinations of each system. Although
several methods are available for studying each system in greater detail, our work has
focused on their application to individuals with various neurological conditions. We
turn now to a review of this work.

4. Perceptual category learning in neurological patients

In the 1980s and 1990s, one of the most successful models of category learning and
recognition memory was exemplar theory [e.g., Nosofsky (1992)]. Exemplar models
assume that people access memory traces (perhaps subconsciously) of exemplars when
asked to recognize or categorize. This theory is parsimonious because it assumes that
the same memory representation underlies both recognition and category learning. If
exemplar theory is correct, people with impaired memory storage or consolidation
processes should show deficits in recognition and category learning [Pickering (1997)].
Amnesic patients have storage and consolidation problems, and generally have damage
to the hippocampus and connected structures (e.g., surrounding medial temporal lobe
regions and the diencephalon), and so these patients should show both types of deficits.
In a classic study, Knowlton, Squire and Gluck (1994) examined probabilistic classifi-
cation learning in a group of amnesic patients. The task (referred to as the weather pre-
diction task) required participants to classify stimuli into one of two categories based on
the relationship (or association) between multiple stimulus attributes. Specifically, par-
ticipants were presented with one to three visual cues and asked to predict whether there
would be “rain” or “sun.” Corrective feedback was provided following each response.
There were 14 different combinations of four cues, and each combination was differen-
tially associated with the probability of “rain” or “sun.” Knowlton, (1994) found that
amnesic patients with damage to the hippocampus or diencephalon performed normally
on this “weather prediction” task (at least for the first 50 trials), whereas these same
patients were impaired when asked specific questions about the learning context (i.e., an
explicit memory task). Thus, amnesic patients were able to learn categories but were
unable to recall consciously the circumstances surrounding their learning, suggesting
that the hippocampus does not mediate early category learning.

It is important to note that amnesiacs showed normal category learning during the first
50 trials, but they did not perform as well as controls later in learning (i.e., during the last
200 trials). Because only 14 unique cue-stimulus combinations were utilized, Knowlton
et al. (1994) suggested that this “late-training deficit” resulted because normal controls
used explicit memory for the stimuli that arose from multiple stimulus presentations,
whereas the amnesic patients were unable to use such information [for a related explana-
tion see Gluck, Oliver and Nyers (1996)]. Recent work also suggests that a number of qual-
itatively different strategies can be used to solve this task, ranging from strategies involving
attention to a single stimulus attribute to the optimal strategy, which involves attention to
all attributes [Gluck, Shohamy and Myers, 2002; see also Ashby and Maddox (2005)].
The Knowlton et al. (1994) study is important because it was one of the first to suggest that category learning and recognition memory might be mediated by different neural substrates. Even so, there are at least two problems with this study. First, the use of only 14 unique cue–stimulus combinations is problematic. This small number of stimuli allows a participant with intact explicit memory to use explicit memory processes to improve categorization performance. To control for the possibility that explicit memory processes will be invoked, and to provide a better test of category learning in amnesia, categories that contain a large number of stimuli should be used. Second, the fact that a large number of qualitatively different strategies can be used to accurately solve the task is problematic. A better approach would be to use a task in which a single, uniquely identifiable, optimal rule (i.e., the rule that maximizes long-run accuracy) can be identified, and for which other strategies yield worse accuracy. Similarly, it would be advantageous to utilize a model-based approach to identify the types of strategies that are being used and to help localize the cognitive processes that lead to any performance decrement.

4.1. Nonlinear information-integration category learning in amnesia

Filoteo, Maddox and Davis (2001a) took such an approach to study category learning in amnesic patients. They utilized the perceptual categorization task [also called the general recognition randomization technique; Ashby and Gott (1988)], which has been used extensively to study category learning in healthy young adults and attentional processes in healthy older adults and patients with PD [see Maddox and Filoteo (in press) for a review]. In a typical perceptual categorization task, participants are presented with simple stimuli, such as a horizontal and a vertical line connected at the upper left corner (see Figure 2), and are asked to assign the stimuli to one of two categories. Prior to the experiment, two bivariate normally distributed categories are specified, and a large number of stimuli are sampled randomly from each bivariate normal distribution. In the Filoteo et al. (2001a) study, 50 unique stimuli were sampled from category A and 50 from category B. With such a large number of unique stimuli, explicit memory processes cannot easily be invoked to improve categorization performance. Because the stimuli are two-dimensional, a unique point in a two-dimensional space can represent each one. Figure 2 depicts the distribution of stimuli used in this study in this two-dimensional space, where the x-axis represents the length of the horizontal line and the y-axis the length of the vertical line. Open squares denote category A stimuli and filled circles denote category B stimuli.

Because the categories are normally distributed, they overlap and a single experimenter-defined categorization rule (i.e., the rule that maximizes long-run accuracy) can be derived [e.g., Maddox and Ashby (1993)]. The form of the rule is determined by the relationship between the two category distributions and thus depends on the relationship between the two stimulus attributes. Filoteo et al. (2001a) examined a complex information-integration categorization rule that was based on a highly nonlinear relationship between the two stimulus attributes. The broken quadratic curve in Figure 2 denotes the optimal categorization rule (or boundary), and yields 95% correct responses.
Because the category structures are defined a priori, the experimenter has a great deal of control over potentially important aspects of the categories, such as the optimal accuracy rate and the shape of the optimal categorization rule (e.g., linear or nonlinear), to name a few. An additional advantage of the perceptual categorization task is that a number of quantitative models of category learning have been developed specifically for application to data collected in this task [Maddox and Filoteo (in press)]. Categorization accuracy (i.e., percent correct) is the typical metric used in neuropsychological studies of category learning, and although its use has several strengths, accuracy analyses have at least two weaknesses. First, because accuracy analyses generally focus on averaged performance (e.g., ANOVA), important individual differences may be obscured. The model-based approach utilized by Filoteo et al. (2001a), on the other hand, allows one to identify and quantify performance at the individual participant level. Second, accuracy-based analyses do not allow the researcher to tease apart the separate effects of various cognitive processes on performance. For example, categorization accuracy is affected not only by participants’ ability to learn the experimenter-defined categorization rule, but also by their ability to accurately apply the learned rule on each trial³. The first process we refer to as categorization rule learning. Difficulty learning the experimenter-defined categorization rule (denoted by the broken curve in Figure 2) will lead to a reduced accuracy level. The second process we refer to as rule application variability. This has to do with participants’ ability to apply consistently from trial to trial whatever categorization rule they might have learned. Greater variability in rule application can also lead to reduced accuracy. Both processes, categorization rule learning and rule application variability, will affect accuracy measures, and thus at the level of accuracy these two processes are nonidentifiable. The model-based approach utilized by Filoteo et al. (2001a) alleviates this problem because it allows one to separate categorization rule learning from rule application variability. The modeling approach will be summarized briefly after we review the experimental findings.

Filoteo et al. (2001a) had two amnesic patients and five matched controls complete six 100-trial blocks of trials in the perceptual categorization task using the Figure 2 category structures. On each trial a stimulus was selected at random and was presented on the computer screen, the participant generated a category A or category B response, and corrective feedback was provided.

The top panel of Figure 3 displays the proportion correct for the amnesic and control participants during the first 100 trials and the final 100 trials (i.e., 501–600) from the first experimental session. One of the amnesic patients and a matched control also completed a second session and the data from the first 100 trials are also presented in Figure 3. Several comments are in order. First, during the first and final blocks from Day 1, the amnesic patients and controls showed equivalent performance. In fact, performance did not differ in any of the six blocks of trials. This finding is important

³ We are using the term “rule” more generally here than in COVIS. In the current application, the “rule” might be verbalizable or nonverbalizable. It might involve learning a decision bound or assigning responses to regions of perceptual space
because it suggests that amnesiacs can learn to categorize, and that the late training deficit observed in the weather prediction task was likely due to the use of explicit memory processes by the control participants. Second, during the first block of trials from Day 2, the amnesic patient and the control again showed equivalent performance, and in fact, performance during the first block of the second session was slightly better than that during the final block of trials from the first session. Some have suggested that amnesic patients learn categorization rules using working or short-term memory processes [Nosofsky and Zaki (1999)]. For example, it has been suggested that amnesic patients are able to take advantage of the repeating stimuli during some categorization tasks and this information is then used to categorize further stimuli [Nosofsky and Zaki (1999)]. The Day 2 results argue against this possibility because it is highly unlikely that subjects were able to make use of working or short-term memory processes between the 2 days. Instead, these findings indicate that the categorization rule was retained over the 1-day delay period, and given the severity of the memory deficit in our amnesic patient (e.g., on Day 2, the patient did not recall having been given the test on the previous day), brain systems not involved in explicit memory likely mediated the retention of this categorization rule.

4.1.1. Model-based analyses

The specifics of the modeling procedure are outlined in numerous articles [e.g., Ashby (1992), Maddox and Filoteo (in press)]. In this section we provide only an overview of the approach, highlighting aspects of the modeling that are relevant to the Filoteo et al. (2001a) study. Filoteo et al. tested amnesic patients and controls in their ability to learn a rule in which correct classification was based on a unique nonlinear (quadratic) relationship between the horizontal and vertical line lengths. This rule is depicted as the broken curve in Figure 2. The aim of the modeling approach with these data was twofold. First, to determine how well a participant learned the optimal decision rule, we fit the optimal decision bound model to each block of data separately for each participant. As a measure of categorization rule learning we examined the goodness-of-fit value (i.e., the

\[ ax^2 + by^2 + cxy + dx + ey + f, \]

where \( a \) to \( f \) denote the coefficients of the quadratic function, and \( x \) and \( y \) denote the horizontal and vertical line lengths, respectively. In the experimenter-defined optimal quadratic categorization rule, the coefficients, \( a \) to \( f \), are fixed, and are determined from the category structures. Maximum likelihood criteria were used to estimate model parameters [see Ashby (1992)]. In essence, the maximum likelihood procedure attempts to maximize the “fit” of the model to the data by attempting to generate predictions from the model that most closely match the observed data. In Filoteo et al. (2001a), the data were the participant’s categorization responses for each presented exemplar. Thus, for each exemplar the observed probability of responding “Category A” was either 1 or 0. Assuming the optimal categorization rule is applied, and for a fixed value of the rule application variance (an estimate of the variability associated with a participant’s inability to accurately apply the same rule on every trial), the model generated a predicted probability of responding “Category A” for each exemplar. Because the coefficients are fixed in the optimal model, the rule application variance was the only parameter adjusted iteratively until the difference between the observed and predicted “Category A” response probabilities was minimized.
maximum likelihood value, -lnL, negative log likelihood) from the optimal model. The smaller the fit, the better the optimal rule describes the data. Second, we were interested in quantifying the magnitude of any variability in the application of the participant’s rule. To achieve this goal we fit a suboptimal model that assumed a quadratic decision bound, but allowed the coefficients of the quadratic decision bound to be estimated from the data. As a measure of rule application variability, we examined the noise variability estimate from this suboptimal model. The smaller the magnitude of the rule application variability, the less variable was the participant’s trial-by-trial application of the rule. To reiterate, at the level of accuracy rates, these very different processes are nonidentifiable. Only with the model-based approach can these two subprocesses be teased apart and be made identifiable.

Because no accuracy differences were observed across amnesic and control participants, Filoteo et al. (2001a) predicted no differences in categorization rule learning or rule application variability. None were observed. Since publication of the Knowlton et al. (1994) and Filoteo et al. (2001a) studies showing intact category learning in amnesia, several challenges to these findings have been offered. For example, Nosofsky and Zaki (1998) suggested that exemplar theory can be used to account for the category learning/recognition memory dissociation observed in amnesia. Smith and Minda (2000) argued against these claims. Although an interesting topic of further work, our laboratories have shifted attention away from amnesia toward a study of category learning in striatal-damaged patients. COVIS does not implicate the medial temporal lobes directly in category learning and so the Filoteo et al. findings were expected. COVIS does predict that the striatum is directly involved in category learning and so an examination of striatal contributions to category learning is in order.

4.2. Nonlinear information-integration category learning in striatal-damaged patients

Knowlton, Mangels and Squire [1996a; see also Knowlton et al. (1996b)] suggested that the striatum may play an important role in category learning. They found that patients with PD, whose neuropathology results in a decrement in striatal functioning, demonstrate impaired probabilistic classification learning in the weather prediction task but intact recognition memory. These observations are also consistent with animal studies that implicate the striatum in certain aspects of category learning [McDonald and White (1993), Packard and McGaugh (1993)]. Notice that these data, along with those from amnesic patients, represent a double dissociation. PD patients show a deficit in category learning but intact recognition memory, whereas amnesic patients show intact category learning but a deficit in recognition memory. These results provide strong evidence for the involvement of the striatum, but not the medial temporal lobes, in information-integration category learning.

Although the Knowlton et al. (1996a) study suggests that the striatum is involved in category learning, other reports argue against this position. One important study was conducted by Reber and Squire (1999), who found that PD patients performed normally in learning to classify dot patterns and artificial grammars. These results contradict the
findings of Knowlton et al. (1996a), who demonstrated that PD patients were impaired in probabilistic classification. In an attempt to reconcile these findings, Reber and Squire (1999) suggested that probabilistic classification is different from artificial grammar or dot pattern classification because the participants must learn the cue-outcome relations through trial-by-trial feedback [for an alternative explanation, see Ashby and Maddox (2005)]. In artificial grammar and dot pattern classification, on the other hand, individuals are simply exposed to members of a category, and are required to study these items. They are then tested on items that either “fit” or “do not fit” the trained category structure. Reber and Squire argue that the need to learn cue-outcome relations in probabilistic classification requires the striatal learning system, which is impaired in PD. Since dopamine is dramatically reduced in the striatum of patients with PD [Cornford, Chang and Miller (1995)], this interpretation would also be consistent with the proposed role of dopamine in reward-based learning mechanisms [Ashby et al. (1998)].

It is very likely that the need to learn cue-outcome relations in probabilistic classification at least partially accounts for the poor performance of PD patients. Unfortunately, the artificial grammar and dot pattern classification tasks differ from the probabilistic classification task in a number of ways, any of which might fully or partially explain the performance dissociation observed in PD. First, the artificial grammar and dot pattern classification tasks usually require the learning of only a single category, whereas two categories must be learned in the probabilistic classification task [Ashby and Maddox (2005)]. Second, perfect performance is possible in the artificial grammar and dot pattern classification tasks, whereas it is not possible in the probabilistic classification task. Third, all three tasks differ with regard to the nature of the stimuli (dot patterns, artificial words, or cards with geometric forms), and the response requirements (point at the center dot, reproduce the artificial word on a piece of paper, choose a category “rain” or “sun”). Finally, the nature of the categorization rule is very different across the three tasks. Dot pattern classification can be solved by prototype extraction, and the artificial grammar task might involve perceptual priming of letter string chunks [Knowlton et al. (1996b)]. The probabilistic classification task, on the other hand, appears to involve the learning of a complex categorization rule [although simpler strategies will suffice; Gluck et al. (2002)]. Thus, the performance dissociation could be due to any one of the following: reliance on striatal learning, effects of maximum attainable accuracy, differential stimulus characteristics and task requirements, categorization rule complexity, or any combination of these factors.

Because of these problems, we decided to examine striatal involvement in category learning using the perceptual categorization task. Maddox and Filoteo (2001) had PD patients and matched controls complete six 100-trial blocks of trials using the nonlinear information-integration category stimuli displayed in Figure 2, and Filoteo, Maddox and Davis (2001b) had HD patients and matched controls perform the same task. Like PD, HD also impacts striatal functioning [Vonsattel et al. (1985)].

The asymptotic accuracy rates obtained during the final block of trials (i.e., trials 501–600) for the PD, HD, and relevant control participants are depicted in the middle panel of Figure 3. Note that both the PD and HD participants showed clear category
learning deficits. To determine the locus of the nonlinear information-integration category learning deficit in PD and HD participants, we examined the categorization rule learning and rule application variability estimates from the final block of trials. These values are displayed in the bottom two panels of Figure 3. The PD patients evidenced categorization rule learning and rule application variability deficits suggesting that their accuracy deficit was due to an inability to learn the optimal rule, and to greater variability in the application of the rule that they had learned. The HD patients showed categorization rule learning deficits but not rule application variability deficits (although the trend is in that direction) suggesting that their performance deficit was due to an inability to learn the optimal rule.

Taken together with the Filoteo et al. (2001a) study, the results support the prediction that the striatum, but not the medial temporal lobes, is involved in nonlinear information-integration category learning when a large number of unique stimuli are utilized to minimize the influence of explicit memory processes. This follows since information-integration category learning is assumed to be mediated within the tail of the caudate, which is impacted in HD, and involves a dopamine-mediated reward signal, which is likely deficient in patients with PD. The results also suggest that the locus of the PD and HD subjects’ deficits was in their ability to learn the optimal decision bound, with the additional difficulty, for PD patients only, of accurately applying the rule that they have learned.

4.3. Rule-based category learning in PD

Maddox and Filoteo (2001) also examined rule-based category learning in PD using the perceptual categorization task and the same two-line stimuli. Their task required the participant to attend to both stimulus dimensions, and to use the following rule: respond A if the vertical line is longer than the horizontal line and respond B if the vertical line is shorter than the horizontal line. They found that PD patients were as good as controls at performing this task.5

This finding is at odds with the predictions of COVIS and of a recent study conducted by Ashby et al. (2003b). COVIS predicts that PD patients will show impaired rule-based category learning because of depleted dopamine projections from the substantia nigra into the head of the caudate nucleus. As a test of this hypothesis, Ashby et al. had PD patients and controls learn a rule-based category structure. The stimuli were constructed from four highly discriminable binary dimensions (e.g., red vs. blue, circle vs. square, etc.), factorially combined to create a set of 16 stimuli. One dimension was chosen to be relevant while the three remaining dimensional values varied randomly across trials. Using a learning criterion of 10 correct responses in a row, Ashby et al. found that ~50% of the PD patients failed to learn this rule-based task within 200 trials, whereas only 10% of the controls failed to learn, suggesting a PD deficit in rule-based category learning.

5 Filoteo et al. (2001b) examined HD patient learning in the same condition and found a small deficit in these patients.
The findings of Ashby et al. (2003b) and Maddox and Filoteo (2001) challenge the simplistic notion that PD patients will always show a rule-based category learning deficit.

Unfortunately, the Maddox and Filoteo (2001) and Ashby et al. (2003b) studies differ in a number of important ways, each of which might explain the contradictory findings. For example, the Maddox and Filoteo study used a large number of overlapping, continuous-valued dimension stimuli that required that a unique decision criterion be learned but did not require that any dimensional information be filtered. On the other hand, the Ashby et al. study used a small number of nonoverlapping, binary-valued dimension stimuli that did not require that a unique decision criterion be learned, but required that variation along three irrelevant dimensions be filtered.

To begin to shed some light on the locus of potential PD deficits in rule-based category learning, Maddox et al. (2005) examined PD patients’ learning in two rule-based category learning conditions. Like Maddox and Filoteo (2001), they used the perceptual categorization task. Thus, they used a large number of overlapping, continuous-valued dimension stimuli that required that a unique decision criterion be learned (similar to that in Figure 1a, but with different stimulus dimensions). However, unlike Maddox and Filoteo (2001), who required attention to both stimulus dimensions, Maddox et al. (2005) required the participant to learn a decision criterion along one stimulus dimension while filtering out (or ignoring) information from a second stimulus dimension. Each participant completed five 50-trial blocks of trials in each condition. For ease of exposition we focus on asymptotic performance collapsed across the two rule-based conditions.

The asymptotic accuracy rates obtained during the final block of trials (i.e., trials 201–250) for the PD and control participants are depicted in the top panel of Figure 4. Notice that the PD patients showed a clear rule-based category learning deficit. To determine the locus of the rule-based category learning deficit in PD, we first applied a series of models to determine whether the PD patients’ deficit was due to an inability to ignore variation along the irrelevant dimension. In other words, were PD patients more likely than controls to use a strategy that was qualitatively different from the optimal strategy – namely an information-integration strategy, when the optimal strategy was to attend selectively to only one dimension? To test this hypothesis, we fit three models. The optimal rule-based model assumes that the participant attended selectively and used the optimal decision criterion. The suboptimal rule-based model assumes that the participant attended selectively, but used a suboptimal decision criterion. The information-integration model assumes that the participant was unable to attend selectively and instead used a linear decision bound constructed from a weighted linear combination of the two dimensional values. The results were clear. For 84% of the PD patients and

---

6 The models are nested in the sense that the suboptimal rule-based model can be derived from the information-integration model by setting the slope of the information-integration model to the optimal slope (zero or infinity), and the optimal rule-based model can be derived from the suboptimal rule-based model by setting the decision criterion to the optimal value. Using nested modeling techniques [see Ashby (1992)], we identified the best-fitting model, defined as the simplest model for which a more general model did not provide a statistically significant improvement.
81% of the control participants, a rule-based model provided the best fit of the data suggesting that the PD deficit was not due to a bias towards integrating information (i.e., using a qualitative different strategy from that of the optimal classifier) when the optimal strategy was to attend selectively.

Fig. 4. Accuracy rates, categorization rule learning, and rule application variability estimates from a rule-based category learning study conducted by Maddox et al. (2005) (see text for details). *Statistically significant performance difference ($p<0.05$).
To further examine the locus of the rule-based category learning deficit in PD, we examined the categorization rule learning and rule application variability estimates from the final block of trials. The categorization rule learning index was determined from the fit of the optimal rule-based model, and the rule application variability index was determined from the noise parameter estimate from the suboptimal rule-based models. These values are displayed in the bottom two panels of Figure 4. The PD patients evidenced categorization rule learning deficits, but not rule application variability deficits (although this difference was marginally significant).

These analyses suggest that PD patients were as likely to select a task-relevant rule as control participants, but their use of this rule was less optimal. One way in which the rule might be used suboptimally is in the placement of the decision criterion. To assess this question, we examined the decision criterion estimates from the suboptimal rule-based model. Specifically, we examined the absolute deviation between the best-fitting decision criterion and the optimal decision criterion. By using the absolute deviation, the analyses assess the extent of a response bias independent of the direction of that bias. For the PD patients, the absolute deviation from optimal was 34 pixels, whereas for controls the absolute deviation was 20 pixels, a reduction of nearly 50%. Thus, PD patients tended to exhibit larger suboptimalities in decision criterion placement than the control participants.

It is worth mentioning that a group of patients with cerebellar lesions (CB) were also tested. CB patients showed no deficits in accuracy, categorization rule learning, or rule application variability, suggesting that the cerebellum is not involved in rule-based category learning.

The Maddox et al. (2005) study offered a straightforward extension of Maddox and Filoteo (2001) to a situation in which irrelevant dimensional information had to be filtered. In a related study, Filoteo et al. (in press) offered a straightforward extension of Ashby et al. (2003b). In Ashby et al. (2003b) 16 four-dimensional, highly discriminable, binary-valued dimension stimuli were used. One dimension was selected as the relevant one, and the other three were irrelevant but varied randomly across trials. Ashby et al. (2003b) found a large performance deficit for PD patients relative to matched controls. Filoteo et al. (in press) were interested in determining whether the number of randomly varying irrelevant dimensions might impact the magnitude of the rule-based category learning deficit in PD. In all conditions, the stimuli were four-dimensional, one dimension was relevant to solving the task, and the other three were irrelevant. However, the number of irrelevant dimensions that could vary randomly was manipulated across conditions. In one condition, all three irrelevant dimension values could vary randomly across trials. This is analogous to the Ashby et al. (2003b) condition. In a second condition, two of the three irrelevant dimension values varied randomly and the third was held fixed across trials. In the third and fourth conditions, one and none of the irrelevant dimension values varied randomly with the remaining ones (two or three, respectively) held fixed across trials. Using the Ashby et al. (2003b) learning criterion of 10 correct responses, Filoteo et al. (in press) found that when no irrelevant dimensional variation occurred, PD patients and controls did not differ in the number of trials it took to learn
the criterion, but as the number of irrelevant dimensions increased, PD patients’ rule-based learning was impacted to a greater extent than that of controls.

Clearly more work is needed to fully understand the properties that affect rule-based category learning in PD. Even so, this collection of studies provides a nice starting point and suggests that the need to filter irrelevant dimension information is predictive of rule-based category learning deficits in PD for both continuous and binary values stimuli.

4.4. Further study of information-integration category learning in PD

Recall that Maddox and Filoteo [2001; see also Filoteo et al. (2001b)] examined PD patients’ ability to learn a nonlinear information-integration categorization rule using the perceptual categorization task. The stimulus was a horizontal and vertical line connected at the upper left corner that varied in length across trials (see Figure 2). A large number of unique stimuli were presented and the categories overlapped. PD patients showed a consistent performance deficit across six 100-trial blocks. Ashby et al. (2003b) also examined information-integration category learning using the same 16 four-dimensional, highly discriminable, binary-valued dimension stimuli used to study rule-based category learning. The only difference was that the stimulus-to-category assignments were modified to construct an information-integration condition. In their information-integration condition, one dimension was irrelevant, and category assignment was based on the combination of information from the three remaining stimulus components. Using a learning criterion of 10 correct responses in a row, Ashby, Noble, et al. found that similar percentages of PD patients and controls (50%) failed to learn the task, suggesting that PD does not result in a deficit in information-integration category learning.

As with the rule-based tasks described above, the results from Maddox and Filoteo (2001) and Ashby et al. (2003b) lead to different conclusions. However, the tasks differ along a number of important dimensions, making it difficult to determine the locus of these contradictory findings. One interesting aspect of the results is the difference in complexity or difficulty of the two tasks. In Ashby et al. (2003b), PD patients and controls who learned the task learned it in 80 trials on average. In Maddox and Filoteo (2001), on the other hand, PD patients showed a deficit relative to the controls across all 600 trials, and never reached 80% accuracy even after 600 trials. These data suggest that the information-integration rule used by Maddox and Filoteo (2001) is more complex and that this complexity might impact the likelihood of observing an information-integration category learning deficit in PD. Even so the tasks differ along too many other dimensions to make any definitive claims.

To further explore the effects of information-integration rule complexity on PD patients’ category learning, Filoteo et al. (2005) tested a group of PD patients and healthy elderly controls using the linear and nonlinear information-integration conditions displayed in Figures 1b and 1c, respectively. The stimulus was a line that varied in length and orientation across trials. Each participant completed six 100-trial blocks of trials in each condition. It is important to reiterate that, because this work is couched within the framework of the perceptual categorization task, a number of important factors are equated
across conditions (e.g., optimal accuracy), while only the form of the optimal decision bound is manipulated.

Before summarizing the results of this study, let us first describe the methods we used to model the data. Since we published our early work on category learning in amnesia, PD, and HD [Filoteo et al. (2001a,b), Maddox and Filoteo (2001)], the focus of our modeling approach has changed slightly. Instead of focusing on estimates of categorization rule learning and rule application variability, we have begun attempting to characterize the strategy that participants are actually using. One conclusion that we have drawn from our parallel work using healthy young adults is that participants often try to solve information-integration tasks using hypothesis-testing strategies when the experimental conditions are not conducive to learning with the procedural-learning system [see Maddox and Ashby (2004) for a review]. This might also occur with PD patients. Since the neurobiological machinery necessary to solve information-integration tasks is deficient in PD, it might be the case that PD patients attempt to use hypothesis-testing strategies. To investigate this possibility, we developed a large number of models that were applied to the data from each block of trials separately for each participant. Some of these models were hypothesis-testing models and some were information-integration models.

Figure 5 displays hypothetical decision bounds and the resulting response regions from specific response strategies that might be applied in the linear information-integration condition. The four models in the leftmost column are hypothesis-testing models and the three on the right are information-integration models. The top two hypothesis-testing models instantiate one-dimensional strategies. One assumes that the participant sets a criterion on length and ignores orientation, whereas the other assumes that the participant sets a criterion on orientation and ignores length. The bottom two hypothesis-testing models instantiate two-dimensional, conjunctive strategies. Each assumes that the participant sets a criterion on the length dimension and a separate criterion on the orientation dimension. In the first case, the participant responds A if the length is “short” and the orientation is “steep,” and B otherwise. In the second case, the participant responds B if the length is “long” and the orientation is “shallow,” and A otherwise. The topmost information-integration strategy assumes that the participant uses the optimal decision bound. The middle model assumes that the participant uses a suboptimal linear decision bound, but allows the slope and intercept to be suboptimal. The bottom model assumes that the participant uses two linear decision bounds.

The hypothesis-testing models applied to the linear information-integration condition (Figure 5) were also applied in the nonlinear information-integration condition.

7 This model is called the Striatal Pattern Classifier [(SPC; Ashby and Waldron (1999)] and was developed as a computational model of the tail of the caudate. The model assumes that there are four “units” in the length-orientation space with two being assigned to category A and two to category B. On each trial the observer determines which unit is closest to the perceptual effect and gives the associated response. The model results in two “minimum-distance-based” decision bounds. This model has been found to provide a good computational model of observers’ response regions in previous information-integration category learning studies [(e.g., Ashby and Waldron (1999), Maddox et al. (2004)].
The information-integration models were identical as well except that the optimal and suboptimal models assumed quadratic bounds.

4.4.1. Brief summary of the results

The asymptotic accuracy rates obtained during the final block of trials (i.e., trials 501–600) for the PD patients and controls are depicted in the top panel of Figure 6. Notice that PD patients showed normal linear information-integration learning, but a deficit in nonlinear information-integration learning. Importantly, PD patients’ deficit in the

Fig. 5. Hypothetical response regions from participants using hypothesis-testing and information-integration strategies to solve the linear information-integration task.
nonlinear condition, but not the linear condition, did not appear to be due to task difficulty per se in that the controls performed somewhat better in the nonlinear condition relative to the linear condition.

The results of the model analyses can also be seen in Figure 6. The middle left panel shows the percent of participants whose final block data were best fit by an information-integration model or a rule-based model. Notice that the model percentages are quite similar across PD and NC participants for the linear and nonlinear information-integration conditions, but many fewer PD and NC participants attempted to use hypothesis-testing strategies in the nonlinear condition. To gain additional insight into the locus of the PD patients’ nonlinear information-integration learning deficit, we
focused only on participants whose data were best fit by an information-integration model. We computed the accuracy rate for these participants (displayed in the middle right panel of Figure 6), as well as the categorization rule learning and rule application variability indices (displayed in the bottom two panels of Figure 6). These analyses suggest that PD patients, who used information-integration strategies to solve the linear information-integration task were as accurate as controls, and showed equivalent rule learning, and equivalent variability in the application of their rule. For PD patients, who used information-integration strategies to solve the nonlinear information-integration task, on the other hand, accuracy was lower, categorization rule learning was poorer, and variability in the application of their rule was higher than that for controls.

Taken together these results suggest that PD patients are relatively normal at solving linear information-integration tasks, but show clear and large deficits in their ability to solve more complex nonlinear information-integration tasks. Although speculative at this point, it seems reasonable to suppose that the “resolution” of the perceptual-decision space in the tail of the caudate must be higher to solve more complex nonlinear information-integration tasks than to solve simpler linear information-integration tasks. It also seems reasonable to suppose that the resolution in the caudate of PD patients is not as high as in controls (possibly due to a weaker dopamine reward signal), leading to difficulty with complex decision rules, but not with simpler ones. Clearly, more work is needed to determine exactly why PD patients are more susceptible to impairment when learning complex information-integration rules.

5. General discussion

The work presented in this chapter builds upon that presented in the previous chapter by Ashby and Valentin. They reviewed the neurobiological underpinnings of a multiple systems model called COVIS. COVIS assumes that learning in rule-based tasks is dominated by an explicit system that uses working memory and executive attention and is mediated primarily by the anterior cingulate, the prefrontal cortex, and the head of the caudate nucleus, whereas learning in information-integration tasks is assumed to be dominated by an implicit procedural-learning-based system, which is mediated largely within the tail of the caudate nucleus, and requires a dopamine-mediated reward signal. Ashby and Valentin reviewed a number of studies conducted using healthy young adults that tested and provided support for several a priori predictions derived from an examination of the neurobiological underpinnings of the two systems.

The current chapter reviewed a body of literature concerning studies conducted in our laboratories that provides a more detailed examination of the systems used by neurological patients with damage to the medial temporal lobes or the striatum. In addition, a quantitative model-based approach was taken in most of these studies that allowed us to tease apart the separate effects of various cognitive processes on performance that are nonidentifiable at the level of accuracy.
The ability of medial temporal lobe amnesic and striatal-damaged patients (PD and HD) to solve a nonlinear information-integration task was examined in three studies [Filoteo et al. (2001a,b); Maddox and Filoteo (2001)]. In each case, a large number of stimuli were presented to alleviate the possibility that participants might recruit explicit memory processes, and participants completed many experimental trials. As predicted by COVIS, medial temporal lobe amnesiacs showed no performance deficit throughout the learning session, whereas both PD and HD patients showed a consistent deficit. Because the optimal nonlinear information-integration rule was unique, we used a model-based approach to determine whether the locus of the accuracy deficit in PD and HD was an inability to learn the rule (categorization rule learning), an inability to apply consistently from trial to trial whatever categorization rule they might have learned (rule application variability), or both. PD and HD patients evidenced deficits in categorization rule learning and rule application variability suggesting that damage to the striatum affects both subprocesses.

The ability of PD patients to solve rule-based category learning tasks was examined across two pairs of related studies [Maddox and Filoteo (2001), Ashby et al. (2003b), Filoteo et al. (2004), Maddox et al. (2005)]. One pair of studies used the perceptual categorization task that utilizes a large number of unique continuous-valued stimuli sampled from overlapping bivariate normally distributed categories. In Maddox and Filoteo (2001), the optimal rule-based strategy required the participant to attend to both stimulus dimensions, whereas in Maddox et al. (2005) the participant was required to attend to only one stimulus dimension while filtering out (or ignoring) information about the second. The two experiments were identical in all other important aspects. When no filtering was required, PD patients performed normally on rule-based category learning, but when filtering was required they showed an accuracy deficit. This accuracy deficit was not due to PD patients’ inability to attend selectively (PD patients used rule-based strategies to the same degree as controls), but rather to the use of a suboptimal decision criterion. These data suggest that PD patients show deficits when the rule-based task requires dimensional filtering.

Unlike the first pair of studies, the second pair of studies utilizes a small number of binary-valued stimuli that were highly discriminable (i.e., no category overlap). Ashby et al. (2003b) used 16 stimuli composed of four binary-valued dimensions. One dimension was relevant to solving the task and the other three were irrelevant. PD patients showed a large rule-based category-learning deficit. Filoteo et al. (in press) utilized similar stimuli and a one-dimensional rule-based category structure, but manipulated the number of irrelevant dimensions, which varied across trials from three [as in Ashby et al. (2003b)] down to zero, with the remaining dimensional values held fixed across trials. Filoteo et al. (2004) found that the magnitude of the PD deficit increased as the number of randomly varying irrelevant dimensions increased. Specifically, when zero or one irrelevant dimension varied, PD patients showed normal rule-based category learning, but when two dimensions varied, PD patients evidenced a rule-based category learning deficit.

The chapter ended with a review of some recent work that examined the effect of the complexity of the information-integration categorization task on PD performance.
Complexity was examined by comparing a linear information-integration rule with a nonlinear information-integration rule. A major focus was also on identifying the types of response strategies that PD patients and controls used to solve these problems. Work with healthy young adults suggests that hypothesis-testing strategies are often used to solve information-integration category-learning problems when the learning situation is non-optimal. Because PD patients have depleted dopamine, it is reasonable to suppose that they might show similar effects. PD patients showed normal category learning when the optimal decision bound was linear, but a deficit in category learning when the bound was nonlinear. The nonlinear information-integration deficit in PD was not due to an increase in the use of hypothesis-testing strategies (nearly the same proportion of PD and control participants used this type of strategy), but rather was due to worse performance by those PD patients using information-integration strategies relative to controls. In summary, the studies reviewed in this chapter represent a first step toward a more detailed understanding of the neurobiological underpinnings of the two category learning systems proposed in COVIS.

References


Chapter 27

NEURAL REGIONS ASSOCIATED WITH CATEGORICAL SPEECH PERCEPTION AND PRODUCTION

SUSAN M. RAVIZZA

University of California, Davis

Contents

Abstract 601
1. Introduction 602
2. Evidence for categorical speech processing 602
3. Prefrontal regions and motor speech categories 605
4. Temporal–parietal regions and acoustic speech categories 609
5. Cerebellar contributions to categorical production and perception 611
6. Concluding remarks 612
References 613

Abstract

Speech perception and production have been characterized as categorical processes. In this paper, I present neuropsychological and neuroimaging evidence that supports the idea that speech processes are categorical as well as discussing what types of representations are important in perception and production tasks. It has become increasingly clear that both prefrontal and temporal–parietal regions are critical for perceiving phonemes categorically and producing phonemes that are good exemplars of a given phonemic category. I suggest that motor speech representations are important for both perception and production and that these are most likely housed in the prefrontal cortex. Moreover, information about acoustic features associated with phonemes almost certainly resides in more posterior cortices. Whereas these cortical regions are thought to maintain motor or acoustic information necessary for successful categorical perception and production, other regions such as the cerebellum can indirectly affect these abilities. Although speech perception and production appear effortless, these abilities rely on intact functioning of a large network of neural regions.
1. Introduction

Speech is not produced with perfect consistency. In the same way that an ice-skater does not perform a routine the same way twice, speakers do not produce words with faultless precision. However, observers can learn to categorize such movements even though they are not produced uniformly. When watching ice-skating, what may appear to be a flurry of indiscriminable movements to us, are discrete movements to judges at the Olympics. Despite the fact that each axel is performed slightly differently than the previous one, judges are able to distinguish axels from other types of jumps such as the Salchow and the Lutz. Although we are born with the ability to distinguish and create speech sounds, we learn through a process of categorization which range of sounds are meaningful in our native language [Norris and Wise (2000), Diehl, Lotto and Holt (2004)]. The prototypes for various speech sounds that we acquire influences how we sort what we hear into phonemes (the smallest units of speech) as well as serving as the ideal that we aim for when speaking. We quickly hone these abilities and, in time, speech production and perception appear effortless.

The ease with which we produce and comprehend speech belies the large number of neural areas supporting speech skills. When these abilities are lost through strokes or neural abnormalities, it becomes clear that categorical perception and production hinge on the intact functioning of several cortical and subcortical areas. With lesions to certain brain regions, people can lose the ability to categorize speech sounds in a meaningful way and to produce speech in a way that allows accurate categorization by others. Neuroimaging studies of speech perception and production also indicate that speech abilities are not under the domain of a single brain area. In this chapter, I will discuss neural regions important for categorical perception and production as well as how these areas work together to support this process. This topic has been explored using both neuropsychological and neuroimaging methodologies, so it is appropriate to include the results of both types of experiments in order to obtain the clearest picture of how speech abilities are realized in the brain.

2. Evidence for categorical speech processing

Speech perception was first characterized as a categorical process in the work of Alvin Liberman [Liberman (1957)] at Haskins Laboratories. With technological improvements that allowed for the creation of intelligible synthetic speech, it became possible to probe speech perception in increasingly precise ways by creating subtle and systematic differences between phonemes. For example, all phonemes are composed of particular patterns of frequency harmonics called formants. By changing the onset/offset frequency or timing of these formants, a continuum of sounds can be made. In Figure 1a, Figure 1a is Plate 27.1a in the Separate Color Plate section, the starting frequency of the second and third formants is gradually changed from a formant pattern that characterizes the syllable /ba/ to that of /da/. Interestingly, listeners do not respond to this gradual change, but instead
identify these speech sounds as one or the other syllable very consistently (see Figure 1b, Figure 1b is Plate 27.1b in the Separate Color Plate section). Even when the task is changed from identification to discrimination, listeners are less accurate at distinguishing phonemes within a category compared to those that lie on either side of the boundary despite the fact that the amount of formant change can be identical in those two conditions. Categorical identification and discrimination of phonemes have been replicated by many researchers employing different types of speech contrasts in a variety of languages [see Diehl et al. (2004) for a review].

Properties of the human auditory system bias the way in which phonemes are classified. For instance, speech events that take place less than 20 ms apart are judged to occur simultaneously rather than serially due to the temporal resolution of our auditory system. This temporal threshold reflects the pattern of categorization of English speakers.
when detecting phonemes that differ in voicing. When voice onset time – the interval between the stop closure and the onset of vocal cord vibration – is under 20 ms, English-speaking listeners will perceive a voiced syllable such as /ba/ whereas syllables with long voice onset times will be perceived as voiceless, as in /pa/. Although phonemic classification schemes may parallel psychophysical properties of human audition, experiential learning modifies which natural boundaries are important in a particular language [Holt, Lotto and Diehl (2004)]. In English, phonemes with voice onset times (VOTs) under 20 ms will be perceived as voiced regardless of whether the voicing precedes or follows the occlusion release. In languages with three voicing contrasts such as Thai, however, listeners have learned to take into consideration the order of voicing and release burst [Gandour and Dardarananda (1984)].

The categorization of speech sounds is thought to rely on the perception of a number of distinctive features, either acoustic or articulatory, that reflect the unique attributes of a particular phoneme (e.g., voicing). Although controversy exists as to which classification scheme most accurately reflects phonemic organization, as well as which phonemic features are most relevant to categorization, there is broad support for the notion that phonemes are classified in terms of critical features [Blumstein (1990)]. Theories suggesting that phoneme classification occurs in terms of these distinctive features are validated by neuropsychological studies of aphasic errors in perception and production. For example, Blumstein and colleagues have demonstrated that substitution errors committed by aphasic patients are more likely to consist of a phoneme that differs in a single feature from the target phoneme [e.g., ball → doll (place of articulation) not ball → call (place of articulation and voicing)] [Blumstein (1973)]. Analogously, aphasic patients are more likely to confuse phonemes that differ along a single feature rather than multiple dimensions [Blumstein, Baker and Goodglass (1977)]. These error patterns provide evidence that phonemes are represented within a structure that makes certain slips of the tongue and misperceptions more likely based on phonological relationships within the classification scheme. If no such categorical structure existed, there should be no systematic bias in phonemic errors [Blumstein (1990)].

Neuroimaging research has also provided evidence for the categorical nature of phonemic processing, although mostly in the realm of perception. A magnetoencephalography (MEG) study of categorical perception [Simos et al. (1998)] demonstrated that different regions of the primary auditory cortex responded to voiced (0–20 ms VOTs) and voiceless phonemes (40–60 ms VOTs). Moreover, these regions did not respond differentially to phonemes within a voicing category. A recent functional magnetic resonance imaging (fMRI) study has implicated nonprimary neural regions in categorical perception as well [Desai et al. (2004)]. Using sine wave stimuli that approximated a place of articulation continuum (/ba/ – /da/), Desai et al. (2004) examined which neural regions were recruited when perception shifted from a continuous to a categorical analysis. Participants performed a discrimination task in a naive state with regard to the potential phonetic properties of the stimuli and then performed the discrimination task again after they were informed that the stimuli could be perceived as a continuum of syllables ranging from /ba/ to /da/. A second continuum of sine wave
tones was also created that did not contain phonemic features; that is, there was no potential for these tones to be perceived as speech. A region of the left superior temporal sulcus was more active when participants were informed of the phonetic properties of the stimuli than when they were categorizing the identical sounds in a naive state. Further, this region was not active in the corresponding nonphonetic condition and its activity was correlated with a behavioral index of categorical perception (greater discrimination accuracy for phonemes across a boundary than within it).

Taken together, both the neuroimaging and neuropsychological literatures substantiate the idea that speech production and perception can be understood as a categorical process. However, as stated previously, it is far from clear what phonemic features are important to such a classification scheme. At a very general level, categorization schemes can vary based on whether they emphasize information about what is heard (acoustic features) versus how sounds are produced (motor features). Most likely, both types of information are important in perception and perhaps production as well. In the following sections, I will discuss how specific neural regions may support categorical perception and production in terms of representing motor or acoustic phonemic information. Then, I will explain how brain regions that do not represent categorical information per se can still affect accurate perceptual classification of phonemes and the ability to produce phonemes that are good prototypes of their phonemic categories.

3. Prefrontal regions and motor speech categories

Traditionally, prefrontal regions, and specifically the opercular cortex (including Broca’s area), have been associated with speech fluency. Patients with lesions to this area [Ravizza (2003)] or who are diagnosed with Broca’s aphasia/apraxia [although note that Broca’s aphasia is not synonymous with damage to Broca’s area; see Ravizza (2001) for a review] are more likely to produce phonemes that are not good exemplars of a given target category. Normally, parameters such as VOT tend to be distributed around the mean for a particular item if a number of productions are measured and, when phonemes are contrastive for a given parameter, these distributions have little overlap (see Figure 2a). Patients who exhibit such characteristics are usually diagnosed with Broca’s aphasia and/or apraxia of speech [see Duffy (1995) for an explanation of the cooccurrence of Broca’s aphasia and apraxia of speech]. Those with lesions to assumed [Gandour and Dardarananda (1984), Blumstein et al. (1977,1980), Baum et al. (1990)] or actual [Ravizza (2003)] prefrontal regions, however, produce features whose distributions tend to overlap either because of increased variability (Figure 2b) or a decreased separation between mean values (Figure 2c). Speech apraxia is not caused by weakness of the articulatory musculature as these patients are often able to correctly produce nonlinguistic orofacial movements [Duffy (1995)]. In conjunction with the fact that apraxic patients are aware of their speech errors [Dronkers (1996)], this type of disorder is thought to be an impairment of motor planning, that is, a correct target phoneme has been selected but the motor attributes associated with production of that phoneme are noisy.
Recently, attempts have been made to find the neural region associated with apraxia of speech by assessing the common area of damage in patients. Dronkers (1996) claimed that the left anterior insula was critical to intact motor speech planning as this region was damaged in all patients with apraxia of speech and spared in all patients without apraxia. Whereas Dronkers (1996) selected chronic patients based on their language diagnosis and then examined where their damage occurred, Hillis et al. (2004) assessed acute patients with or without lesions to a region of interest and then explored whether apraxia could be predicted from damage to a particular region. In contrast to Dronkers (1996), Hillis et al. (2004) found no relationship between apraxia of speech and insular damage, but instead demonstrated that damage to Broca’s area was critical to motor planning deficits. The divergent findings in these two studies reveal the complexity of categorical speech production [Dronkers and Ogar (2004)] and indicate the need for a better understanding of how separate regions of the prefrontal cortex support speech fluency. For example, Broca’s

---

**Fig. 2.** Illustration of the production of speech tokens along a given parameter (e.g., voice onset time). Typically, two contrastive phonemes such as “ba” and “pa” show (a) little overlap in their distributions, however, brain damage can induce (b) more variability in production or (c) a reduced separation of mean values resulting in tokens that are less prototypical of their phonemic category.
area and the insula may provide unique contributions to speech production with the former area more important in top–down control of speech planning [e.g., when speech is effortful or items are novel; Chein, Ravizza and Fiez (2003)] and the latter area more involved in routine speech production, especially if it is performed overtly [Ackermann and Riecker (2004)]. Moreover, it may be that functions associated with Broca’s area are more readily taken over by its right-sided homologue or other motor regions not specialized for speech such as the supplementary motor area (SMA) while the functions of the insula are more circumscribed. In both cases, the duration of the recovery interval at the time of testing could influence which region was most implicated in categorical speech production.

Given that the primary motor cortex is situated within the frontal lobe, it is not surprising that nearby prefrontal regions such as Broca’s area are involved in motor speech planning and execution. More surprising to some is the mounting evidence that prefrontal regions are also involved in categorical perception. Broca’s aphasics/apraxics [see Blumstein (1990) for a review] and frontal patients [Ravizza (2003)] often show difficulty in identifying and discriminating phonemes and display more continuous, rather than categorical, functions (see Figure 3). Neuroimaging studies have also shown that prefrontal regions recruited in speech production are also active in listening to stories [Papathanassiou et al. (2000)]. In addition, a recent study by Watkins and Paus (2004) demonstrated increased excitability of the orofacial muscles during auditory speech perception, which also was related to activity of Broca’s area. Thus, neural regions associated with speech production are also recruited when perceiving speech.

These findings are consistent with motor theories of speech perception [Liberman (1993)] that emphasize the importance of articulatory information to this ability [see Diehl et al. (2004) for a review]. According to this class of theories, phonemes are categorized based on how they are produced. It is possible, then, that prefrontal regions maintain abstract motor representations that are used for speaking as well as comprehending. However, the results of neuroimaging studies have provided only mixed support for the role of speech production regions in perception. Studies examining neural regions recruited in passive listening tasks have not consistently reported Broca’s area and the left premotor regions to be among these regions [Petersen et al. (1988), Zatorre (1992), Fiez et al. (1996)]. One possible resolution of these discrepant results may be that motor information is redundant with more acoustic representations, and so is only used when acoustic information does not lead to a clear categorical decision. If resolution of the speech signal is not difficult or does not need to be attended to, further processing in terms of motor representations may be aborted. One prediction from this claim would be that the time of recruitment for perceptual regions should differ; that is, prefrontal regions should be recruited after more posterior areas. In fact, an imaging study using a phoneme monitoring task reported that the peak response in Broca’s area occurred after those in the temporal–parietal cortices [Thierry et al. (1999)].

Alternatively, the inconsistent engagement of Broca’s area in passive listening may indicate that this region is not important for housing motor-based phonological representations. Instead, it may be important for attention-based retrieval. For example, in both production and perception, Broca’s area tends to be recruited when the task
becomes more difficult, as when speaking is effortful [see Mechelli and Gorno-Tempini et al. (2003) for a review of areas involved in nonword reading] or in perceptual tasks where attention demands are high [Zatorre et al. (1992), Burton, Small and Blumstein (2000)]. Norris and Wise (2000, p. 873) argue that,

If we find that a particular brain area activates only when performing an explicit metalinguistic task, we have no evidence that this area is directly involved in the normal phonetic or phonological processing of speech. The area could be responsible solely for interrogating the normal speech recognition systems in order to generate a response in this particular task.

Fig. 3. Accuracy of healthy control participants, cerebellar, Parkinson, prefrontal, and temporoparietal patients when identifying items along syllable continuums ranging from (a) “ba”–“da” and (b) “ba”–“pa”. These data were presented in Ravizza (2003).
Several imaging studies have reported greater Broca’s area activity in more attention-demanding perceptual tasks [Zatorre et al. (1992), Burton et al. (2000)]. For example, Myers and Blumstein (2004) reported that activity in the left inferior frontal gyrus was greater for more ambiguous items on a VOT continuum than for end-point items. Thus, it may be that prefrontal regions are necessary only for controlled tasks when conscious access to phonological information is necessary.

In most experiments, the need for motor information and the difficulty of the task are confounded; that is, more difficult phonological tasks may place heavier demands on the motor speech system. However, there are a few lines of evidence that suggest Broca’s area is involved in the representation of articulatory information per se and not just in the controlled retrieval of such information. First, experiments showing increased excitability of the motor system during speech perception and its relationship to Broca’s area activity have used passive listening tasks with low attentional demands [Watkins, Strafella and Paus (2003), Watkins and Paus (2004)]. Second, Broca’s area is important in the observation/imitation of movements in general. Imaging studies have reported activity in Broca’s area when preparing to copy hand movements [Krams et al. (1998)] and in passively viewing movements [Rizzolatti et al. (1996), Iacoboni et al. (1999)]. Moreover, regions involved in speech production in a verb generation task were also active when viewing pictures of movement [Hamzei et al. (2003)]. The fact that Broca’s area is engaged in the observation and imitation of all types of movement argues for a motor contribution of this region to speech tasks rather than retrieval or cognitive control. Third, Broca’s aphasics do not show normal priming effects, as would be expected if only controlled processes were impaired. In a recent experiment, Broca’s aphasics (most of whom had prefrontal lesions) were no faster in a lexical decision task when identical words (e.g., take → take) were presented, and they were inhibited with rhyme primes (e.g., take → sake) [but see Gordon and Baum (1994), Blumstein (2000)]. As rhyme judgments are thought to rely more heavily on articulatory information [Paulesu, Frith and Frackowiak (1993)], the lack of priming effects in Broca’s aphasics may indicate disrupted or noisy motor speech representations.

While it is clear that prefrontal regions support categorical speech production and perception, many questions remain as to their specific contribution. With regard to production, the unique function of areas such as the premotor cortex, the SMA, Broca’s area, and the insula is quite vague. Although there is increasing support for prefrontal regions (especially Broca’s area) being involved in representing motor information used to classify phonemes as well as in producing good phonemic exemplars, explanations in terms of cognitive control rather than articulatory processing per se have not been ruled out.

4. Temporal–parietal regions and acoustic speech categories

Similar to conjectures of a dual role for frontal regions in both speech production and perception, temporal–parietal regions are also thought to contribute to both tasks. The importance of the posterior regions for speech perception has been well documented in both neuropsychological [see Ravizza (2001) for a review] and neuroimaging studies
For example, poor speech comprehension is typically associated with posterior lesions in the vicinity of the Sylvian fissure, although a profound loss requires bilateral lesions [Hickok (2000)]. As can be seen in Figures 3a and b, patients with temporal–parietal lesions do not show the normal categorization function when identifying phonemes [Ravizza (2003)]. Imaging studies have also reported temporal regions to be more active in categorical speech perception tasks than for perceptual tasks of tonal stimuli that are perceived more continuously [Poeppel et al. (2004)]. Unlike prefrontal regions, temporal–parietal sites are active in passive listening to speech [Petersen et al. (1988), Zatorre (1992), Fiez et al. (1996), Crinion et al. (2003)] and in categorical perception regardless of discrimination difficulty [Burton et al. (2000)].

It has become increasingly clear, however, that posterior regions are also important for intact categorical production. Neuroimaging studies of speech production report activity in temporal–parietal regions even when speech is produced covertly [and so the auditory regions are not receiving any overt speech signal; Ackermann et al. (1998), Wise et al. (2001)]. Studies of Wernicke’s aphasics [see Blumstein (1990) for a review] and temporal–parietal patients [Ravizza (2003)] also suggest that lesions to posterior areas result in disrupted categorical production, although these impairments appear to be less severe than those caused by frontal damage. Indeed, some have postulated that, compared to Broca’s aphasics, Wernicke’s aphasics are less likely to produce phonemes that are ambiguous and, instead, produce good exemplars of a contrastive phonemic category [Itoh et al. (1982)].

Note that I have used the term “acoustic” rather than “phonological” to describe the type of information maintained in the temporal–parietal regions. Several lines of research have shown that areas responsible for acoustic analysis are also concerned with extracting phonemic features. Thus, categorical perception can occur with non-speech [Simos et al. (1998)] and regions of the temporal–parietal area are active in the perception of such stimuli [Papanicolaou et al. (2003)]. Indeed, Saygin et al. (2003) demonstrated that the location of the lesion for patients who were impaired at speech and nonverbal sound identification was in the same region of the temporal–parietal cortex.

The more mild impairment of temporal–parietal patients in speech production as well as the fact that errors tend to involve pure substitution rather than poorly articulated items, suggests that this region makes a nonmotor contribution to speech. Temporal–parietal regions are potentially important for the selection of the correct phonological features in production and the failure to associate these features with phonological categories in perception. Again, which features are important in phonological categorization – motor or acoustic – is a controversial issue, but several lines of research have demonstrated the importance of acoustic information in categorizing phonemes [Klatt (1979), Stevens and Blumstein (1981)]. Given that similar regions are involved in both speech and nonspeech perception [Saygin et al. (2003)], auditory association areas in the temporal–parietal area are the most probable sites for the maintenance of such acoustic features.
5. Cerebellar contributions to categorical production and perception

Many neural regions support the ability to produce and perceive speech in a categorical fashion (e.g., cerebellum, basal ganglia, thalamus). These regions are not thought to house the phonemic representations used in the categorization of speech sounds but, instead, affect the accuracy of the information on which categorical judgments are based or the implementation of motor commands. Cerebellar contributions to speech tasks are a good case in point of the impact of noncategorical processes on the efficacy of categorical perception and production undertaken by the frontal and temporal–parietal cortices. Moreover, cerebellar findings in studies of phonological processing illustrate how the pattern of involvement differs between regions thought to be directly involved in phonemic categorization those that have only an indirect effect. For example, whereas it is often difficult to dissociate the performance of frontal and temporal–parietal patients in various speech tasks [Blumstein (1990)], cerebellar patients show a different pattern of results from cortical patients.

The cerebellum has traditionally been acknowledged as a structure that is important for precise motor coordination, and speech production impairments of cerebellar patients have been well-documented [Kent, Netsell and Abbs (1979)]. Lesions to the medial superior cerebellum are associated with ataxic dysarthria [Ackermann et al. (1992)], an inability to control the muscles of articulation where both speech and non-speech movements are disrupted [Dronkers, Redfern and Knight (2000)]. In phonemic production tasks, cerebellar patients often show increased variability but a clear separation of the mean values (e.g., VOT) of a given feature [see Figure 2b; Gandour and Dardarananda (1984), Ivry and Gopal (1993), but see also Ackermann and Hertrich (1997), Ackermann et al. (1999), Ravizza (2003)]. Some authors have suggested that these findings are consistent with a general role of the cerebellum in temporal computations of short duration [<1 s; Ivry and Gopal (1993), Ackermann and Hertrich (1997)]. Thus, whereas both frontal and temporal–parietal patients can show both increased variability and reduced mean values, cerebellar patients tend to show only increased variability [Ravizza (2001)]. Moreover, cerebellar patients are more impaired at implementing only certain phonemic features whereas cortical patients demonstrate more global deficits. For example, in cases where explicit timing is required, cerebellar patients are particularly impaired. Studies have demonstrated that cerebellar patients produce more variable VOTs for voiceless phonemes such as /p/, but are much less impaired at producing voiced cognates (e.g., /b/) [Ackermann and Hertrich (1997), Ravizza (2003)]; the production of voiceless sounds has also been argued to require explicit timing whereas the short VOTs associated with voiced phonemes may be a by-product of the sequence of articulatory movements [Cooper (1977)]. However, prefrontal patients tend to be impaired at producing both voiced and voiceless phonemes [Ravizza (2003)].

Analogous to their production deficits, cerebellar patients do not show categorical perception of phonemes that requires the derivation of temporal parameters. Using a phonemic continuum where the only cue to voicing was the length of a silent interval,
Ackermann and colleagues have demonstrated impairments in categorical perception with cerebellar patients [Ackermann et al. (1998)]. Moreover, this finding has been replicated in neuroimaging studies [Burton et al. (2000), Mathiak et al. (2002)], that is, the right cerebellar hemisphere is more active in conditions where temporal cues are more heavily weighted or when spectral cues are unavailable. These results are in line with the results of other studies demonstrating cerebellar involvement in nonverbal perceptual tasks (e.g., velocity judgments of moving nonverbal stimuli) that nonetheless require temporal processing [Ivry and Diener (1991)]. Again, cortical patients tend to exhibit perceptual problems on many different phonemic tasks regardless of whether temporal or spectral features are important for categorization.

In sum, the cerebellum is engaged on only a subset of the tasks that recruit the frontal and temporal–parietal regions. This accords well with the idea that these cortical regions are important in maintaining the phonological representations used in categorical perception and production, whereas the cerebellum affects these processes indirectly.

6. Concluding remarks

There are many neural regions associated with categorical perception and production, and a discussion of all of them is beyond the scope of this paper. Even within regions of the prefrontal and temporal–parietal cortices, the designation of unique functions to separate areas is not well understood. To make matters more complicated, many functions have been attributed to the same neural region. For example, the parietal operculum has been claimed to be involved in auditory-motor integration [Wise et al. (2001)], auditory imagery [McGuire et al. (1996)], visuomotor speech perception [Calvert and Campbell (2003)], phonological perception [Caplan, Gow and Makris (1995)], and verbal working memory [Paulesu et al. (1993)]. Clearly, the processes needed for producing and perceiving speech are quite complex; associating these processes with the function of various neural structures is a daunting task. However, neuroimaging and neuropsychological research has provided valuable insight in characterizing the types of processes and information important for speech. Indeed, novel evidence for hypotheses regarding the relevance of articulatory information in categorical perception has been provided through careful examination of frontal patients’ performance and the activity of motor regions in imaging studies of perceptual tasks.

I have suggested that, based on the evidence to date, motor representations used for both categorical production and perception are instantiated in Broca’s area. Other regions in the prefrontal cortex such as the SMA and the insula have been shown to be important in categorical production, but it is unclear whether these areas are also involved in perception and how they differ from Broca’s area in terms of their specific contributions to speech production. Also critically important for categorical speech processing are temporal–parietal regions. Although various subregions with this area make unique contributions to speech, their functions are most likely to be consistent with the representation of phonological information based on acoustic information. Finally, I
have discussed how a region not thought to be involved in categorical processing *per se* indirectly affects the ability to produce good exemplars and classify phonemes in some situations. As with any ability, a large network of neural regions contributes to successful speaking and comprehending, contradicting the apparent ease with which these tasks are accomplished.

References


PART 6

CATEGORIES IN PERCEPTION AND INFERENCE
This page intentionally left blank
Chapter 28

SITUATED CONCEPTUALIZATION

LAWRENCE W. BARSALOU

Emory University

Contents

Abstract 620
1. Introduction 621
   1.1. Conceptual systems 621
   1.2. Semantic memory 621
2. Grounding the conceptual system in the modalities 622
   2.1. Modal reenactments of perception, action, and introspection 623
      2.1.1. Storage of modality-specific states that arise in feature systems 623
      2.1.2. Reenactments of modality-specific states 623
   2.2. Simulators and simulations 624
      2.2.1. Simulators 624
      2.2.2. Simulations 625
      2.2.3. Sources of simulators 625
   2.3. Situated conceptualizations 626
      2.3.1. Multimodal simulations implement situated conceptualizations 626
      2.3.2. Entrenched situated conceptualizations 627
   2.4. Inference via pattern completion 628
      2.4.1. Pattern completion with entrenched situated conceptualizations 628
      2.4.2. The statistical character of inference 629
3. Empirical evidence 629
   3.1. Behavioral evidence for a modal nonmodular conceptual system 629
      3.1.1. Predictions for modular amodal vs. nonmodular modal theories 630
      3.1.2. Assessing the presence of modality-specific effects in conceptual processing 631
      3.1.3. Occlusion during property generation 631
      3.1.4. Size during property verification 632
      3.1.5. Shape during property verification 633
      3.1.6. Modality switching during property verification 634
      3.1.7. Shape and orientation during comprehension 635
      3.1.8. Movement direction in comprehension 635
      3.1.9. Further evidence for simulation from comprehension studies 636
      3.1.10. Behavioral evidence for embodiment in social cognition 636
Abstract

Two themes about the conceptual system are developed: (1) modal simulations underlie conceptual processing; (2) conceptual representations are situated. The construct of situated conceptualization—a multimodal simulation that supports one specific course of situated action with a particular category instance—integrates these themes. A given concept produces many different situated conceptualizations, each tailored to different instances in different settings. A situated conceptualization creates the experience of “being there” with a category instance in a setting via integrated simulations of objects, settings, actions, and introspections. On recognizing a familiar type of instance, an entrenched situated conceptualization associated with it becomes active, which provides relevant inferences via pattern completion. Supporting empirical evidence from cognitive psychology, social psychology, and cognitive neuroscience is reviewed.
1. Introduction

1.1. Conceptual systems

The conceptual system is a system distributed throughout the brain that represents knowledge about the world. It is not a collection of holistic images like those in a camera, video recorder, or audio recorder. Instead, it is a collection of category representations, with each category corresponding to a component of experience—not to an entire holistic experience. Across a person’s life span, category knowledge about such components develops for objects, locations, times, events, introspective states, relations, roles, properties, etc.

The conceptual system provides representational support across the spectrum of cognitive activities. In online processing, as people pursue goals in the environment, the conceptual system contributes in several important ways. First, it helps construct perceptions through figure-ground segregation, anticipation, and filling in. Second, it predicts entities and events likely to be perceived, speeding their processing. Third, it supports categorization, assigning perceived entities and events to categories. Fourth, it provides inferences following categorization that constitute expertise about the world. Rather than starting from scratch when interacting with an entity or event, agents benefit from knowledge of previous category members.

The conceptual system is also central to offline processing when people represent non-present entities and events in memory, language, and thought. In memory, the conceptual system provides elaboration at encoding, organizational structure in storage, and reconstructive inference during retrieval. In language, the conceptual system contributes to the meanings of words, phrases, sentences, and texts, and to the inferences that go beyond them. In thought, the conceptual system provides representations of the objects and events that are the objects of reasoning, decision making, and problem solving.

1.2. Semantic memory

Since the cognitive revolution, theorists have proposed many accounts of the conceptual system. The dominant theory, however, has been the semantic memory view, which arises from a proposed distinction between semantic and episodic memory [Tulving (1972)]. Specific models that instantiate this view include network models [e.g., Collins and Quillian (1969), Collins and Loftus (1975)] and feature set models [e.g., Rosch and Mervis (1975), Hampton (1979)]. For a review of semantic memory models, see Smith (1978). This approach to thinking about the conceptual system remains dominant. Researchers throughout the cognitive sciences continue to adopt various forms of semantic memory models in their working accounts of the cognitive system.

Four assumptions underlie the semantic memory view. First, semantic memory is viewed as a modular system, that is, as being autonomous relative to the episodic memory system and to the systems for perception, action, emotion, and motivation. From this theoretical perspective, the conceptual system does not share representation and
processing mechanisms with these other brain systems, but is an independent module that operates according to different principles.

Second, and relatedly, semantic memory representations are assumed to be amodal, differing significantly from representations in modality-specific systems. Specifically, semantic memory representations are assumed to be redescriptions or transductions of modality-specific representations into a new representation language that does not have modality-specific qualities. Instead, these representations consist of arbitrary amodal symbols that stand for modality-specific states and for the entities in the world that these states represent.

Third, semantic memory representations are viewed as decontextualized. In the typical theory, the representation of a category is a prototype or definition that distills relatively invariant properties across exemplars. Lost in the distillation are idiosyncratic properties of exemplars and background situations. Thus the representation of [CAT] might be a decontextualized prototype that includes CLAWS, WHISKERS, and TAIL, with idiosyncratic properties and background situations filtered out\(^1\). As a result, category representations in the semantic memory view have the flavor of encyclopedia descriptions in a database of categorical knowledge about the world.

Fourth, semantic memory representations are typically viewed as being relatively stable. For a given category, different people share roughly the same knowledge, and the same person uses the same knowledge on different occasions.

2. Grounding the conceptual system in the modalities

A diametrically opposed way of thinking about the conceptual system is developed here. The first section of this chapter presents the theoretical assumptions of this approach; the second section presents empirical support for it.

In the theoretical section, the first of three subsections introduces the constructs of reenactment, simulator, and simulation. Rather than being modular, the conceptual system shares fundamental mechanisms with modality-specific systems. As a result, conceptual representations are modal, not amodal.

The second subsection introduces the construct of situated conceptualization, which grounds conceptual processing in situated action. Rather than being decontextualized and stable, conceptual representations are contextualized dynamically to support diverse courses of goal pursuit.

The third subsection illustrates how situated conceptualizations support conceptual inferences via pattern completion. When one part of a situated conceptualization is perceived, the remainder of the conceptualization becomes active, constituting inferences about the current situation.

---

\(^1\) Following the conventions used throughout the Handbook, uppercase will be used to indicate conceptual representations, and italics will be used to indicate linguistic forms. Within conceptual representations, uppercase in brackets will indicate categories, whereas uppercase font with no brackets will indicate properties of categories. Thus, [CATS] indicates a category, whereas CLAWS indicates a property, with cats and claws indicating the respective linguistic forms.
2.1. Modal reenactments of perception, action, and introspection

The modal reenactment of perceptual, motor, and introspective states constitutes the central mechanism in this approach, not amodal redescriptions of these states [e.g., Damasio (1989), Barsalou (1999b, 2003a), Simmons and Barsalou (2003)]. The reenactment process underlying knowledge is assumed to be approximately the same as the reenactment process underlying mental imagery [e.g., Barsalou (1982), Finke (1989), Kosslyn (1994), Zatorre (1996), Farah (2000), Grezes and Decety et al (2001)]. The reenactment process has two phases: (1) the storage of modality-specific states, and (2) the partial reenactment of these states. Each phase is addressed in turn.

2.1.1. Storage of modality-specific states that arise in feature systems

When a physical entity is experienced, it activates feature detectors in the relevant brain systems. During visual processing of a cat, for example, neurons fire for edges and planar surfaces, whereas others fire for color, configural properties, and movement. The overall pattern of activation across this hierarchically organized distributed system represents the entity in vision [e.g., Zeki (1993), Palmer (1999)]. Analogous patterns of activation in other sensory modalities represent how the cat might sound and feel. Activations in the motor system represent actions on the cat. Similar mechanisms underlie the introspective states that arise while interacting with an entity. For example, activations in the amygdala and orbitofrontal areas might represent emotional reactions to the cat.

When a pattern becomes active in a feature system, conjunctive neurons in association areas capture the pattern for later cognitive use. A population of conjunctive neurons codes the pattern, with each individual neuron participating in the coding of many different patterns (i.e., coarse coding). Damasio (1989) calls these association areas convergence zones, and proposes that they exist at multiple hierarchical levels in the brain [also see Simmons and Barsalou (2003)]. Locally, convergence zones near a modality capture activation patterns within it. Association areas near the visual system capture patterns there, whereas association areas near the auditory system capture patterns there. Downstream in more anterior regions, higher association areas in the temporal, parietal, and frontal lobes integrate activation across modalities.

2.1.2. Reenactments of modality-specific states

The convergence-zone architecture has the functional ability to produce modality-specific reenactments. Once a set of conjunctive neurons captures a feature pattern, this set can later activate the pattern in the absence of bottom-up stimulation. When retrieving a memory of a cat, conjunctive neurons partially reactivate the visual state active during its earlier perception. Similarly, when retrieving an action performed on the cat, conjunctive neurons partially reactivate the motor state that produced it. A reenactment never constitutes a complete reinstatement of the original modality-specific state.
Furthermore, bias may often distort it. Thus, a reenactment is always partial and potentially inaccurate. Nevertheless, some semblance of the original state is reactivated—not an amodal redescription.

The reenactment process is not necessarily conscious. Although conscious reenactment is viewed widely as the process that underlies mental imagery, reenactments need not always reach awareness. Unconscious reenactments may often underlie memory, conceptualization, comprehension, and reasoning [Barsalou (1999b)]. Although explicit attempts to construct mental imagery may create vivid reenactments, many other cognitive processes may rely on less conscious reenactments, or reenactments that are largely unconscious.

2.2. Simulators and simulations

Barsalou (1999b, 2003a) developed a theory of the conceptual system based on the neural reenactment of modality-specific states. According to this view, a fully functional conceptual system can be built on reenactment mechanisms. Using these mechanisms, it is possible to implement the type-token distinction, categorical inference, productivity, propositions, and abstract concepts. Contrary to previous arguments, amodal symbols are not the only possible way to implement these classical functions of a conceptual system.

Simulators and simulations constitute the two central constructs of this theory. Simulators integrate information across a category’s instances, whereas simulations are specific conceptualizations of the category. Each is addressed in turn.

2.2.1. Simulators

Categories tend to have statistically correlated properties [e.g., McRae, de Sa and Seidenberg (1997)]. Thus, encountering different instances of the same category should tend to activate similar neural patterns in feature systems [e.g., Farah and McClelland (1991), Cree and McRae (2003)]. Furthermore, similar populations of conjunctive neurons in convergence zones – tuned to these particular conjunctions of features – should tend to capture these similar patterns [Damasio (1989), Simmons and Barsalou (2003)]. Across experiences of instances and settings, this population of conjunctive neurons integrates modality-specific properties, establishing a multimodal representation of the category. Barsalou (1999b) refers to these distributed systems as simulators. Conceptually, a simulator functions as a type. It integrates the multimodal content of a category across instances, and provides the ability to interpret later individuals as tokens of the type [Barsalou (2003a)].

Consider the simulator for the category of [CATS]. Across learning, visual information about how cats look becomes integrated in the simulator, along with auditory information about how they sound, somatosensory information about how they feel, motor programs for interacting with them, emotional responses to experiencing them, and so forth. The result is a distributed system throughout the brain’s feature and association areas that accumulates conceptual content for the category.
2.2.2. Simulations

Once a simulator becomes established for a category, it reenacts small subsets of its content as specific simulations. All of the content in a simulator never becomes active simultaneously. Instead, only a small subset becomes active to represent the category on a particular occasion [e.g., Barsalou (1987, 1989, 1993)]. For example, the [CAT] simulator might simulate a sleeping kitten on one occasion, whereas on other occasions it might simulate a hissing tom cat or a purring house cat. Because all the experienced content for cats resides implicitly in the [CATS] simulator, many different subsets can be reenacted on different occasions.

Simulations serve a wide variety of cognitive functions. As Barsalou (1999b, 2003a) illustrates, simulations can represent a category’s instances in their absence during memory, language, and thought. Simulations can be used to draw inferences about a category’s perceived instances using the pattern completion described later. They can be combined productively to produce infinite conceptual combinations. They also can represent the propositions that underlie type-token predication and recursion.

Simulations can also be used to represent novel category instances not already stored in a simulator. Instances stored on previous occasions may merge together at retrieval, thereby producing reconstructive and averaging effects. Remembering a cat seen once, for example, may be distorted toward a similar cat seen many times. Furthermore, intentional attempts to combine simulations of conceptual components can produce simulations never experienced. For example, people can simulate a cat and then systematically vary simulations of its color and patterning to represent a wide variety of novel instances.

2.2.3. Sources of simulators

In principle, an infinite number of simulators can develop in memory for all forms of knowledge, including objects, properties, settings, events, actions, introspections, and so forth. Specifically, a simulator develops for any component of experience that attention selects repeatedly [Barsalou (1999b, 2003a)]. When attention focuses repeatedly on a type of object in experience, such as for [CATS], a simulator develops for it. Analogously, if attention focuses on a type of action ([BRUSHING]) or on a type of introspection ([HAPPINESS]), simulators develop to represent them as well. Such flexibility is consistent with the Schyns, Goldstone and Thibaut (1998) proposal that the cognitive system acquires new properties as they become relevant for categorization. Because selective attention is flexible and open-ended, a simulator develops for any component of experience selected repeatedly.

A key issue concerns why attention focuses on some components but not on others, such that simulators develop for those components. Many factors influence this process, including genetics, language development, culture, and goal achievement. A further account of these mechanisms lies beyond the scope of this chapter. Notably, though, this is the classic problem of what constrains knowledge (e.g., Murphy and Medin (1985)]. Any theory – not just this one – must resolve it.
Another key issue concerns how simulators for abstract concepts are represented. Barsalou (1999b) proposed that simulators for abstract concepts generally capture complex multimodal simulations of temporally extended situations, with simulated introspective states being central. Relative to concrete concepts, abstract concepts tend to contain more situational and introspective information than do concrete concepts. One sense of [TRUTH], for example, begins with a speaker making a claim about a situation, such as “It’s sunny outside.” A listener then represents the claim, compares it to the actual situation, and decides if the claim interprets the situation accurately. This sense of [TRUTH] can be represented as a simulation of the situation, including the relevant introspective states (e.g., representing, comparing, deciding). Many abstract concepts, such as [FREE-DOM] and [INVENT], can similarly be viewed as complex simulations of situations, with simulated introspective states being central. Wiemer-Hastings, King and Xu (2001) and Barsalou and Wiemer-Hastings (2005) offer preliminary support for this proposal.

2.3. Situated conceptualizations

Barsalou (2003b) contrasts two ways of thinking about concepts [also see Barsalou (1999a)]. On the one hand, semantic memory theories implicitly view concepts as detached databases. When a category is learned, its properties and exemplars are integrated into a general description that is relatively detached from the goals of specific agents. On different occasions, a person uses the same general description to represent the category. Alternatively, a concept can be viewed as an agent-dependent instruction manual that delivers specialized packages of inferences to guide an agent’s interactions with particular category members in specific situations. Across different situations, a concept delivers different packages of inferences, each tailored to current goals and constraints. Because a single general description would be too vague to support all the relevant inferences in a particular situation, more specialized representations are constructed instead.

Barsalou (2003b) referred to one particular package of situation-specific inferences as a situated conceptualization. Consider the concept of [CAT]. According to traditional views, [CAT] is represented as a detached collection of amodal facts that becomes active as a whole every time the category is processed. Alternatively, a simulator for [CAT] produces many different situated conceptualizations, each tailored to helping an agent interact with cats in a different context – no general description of the category exists. For example, one situated conceptualization for [CAT] might support interacting with a playful kitten, whereas others might support interacting with a mean tom cat, or with a purring house cat. In this view, the concept for [CAT] is not a detached global description of the category. Instead, the concept is the skill or ability to produce a wide variety of situated conceptualizations that support goal achievement in specific contexts.

2.3.1. Multimodal simulations implement situated conceptualizations

Barsalou (2003b) further proposed that a complex simulation becomes active across modalities to implement a situated conceptualization. Consider a situated
conceptualization for interacting with a purring house cat. This conceptualization is likely to simulate how the cat might appear perceptually. When cats are purring, their bodies take particular shapes, they execute certain actions, and they make distinctive sounds. All these perceptual aspects can be represented as modal simulations in the situated conceptualization. Rather than amodal redescriptions representing these perceptions, simulations represent them in the relevant modality-specific systems.

A situated conceptualization about a purring house cat is likely to simulate actions that the agent could take in the situation, such as scratching the cat. Modal simulations can also represent these aspects of a situated conceptualization via simulations of the actions themselves, not by amodal redescriptions.

A situated conceptualization about a purring house cat is likely to include simulations of introspective states. Because people experience particular introspections around purring house cats, the respective situated conceptualizations include simulations of emotions, evaluations, motivations, cognitive operations, etc.

Finally, a situated conceptualization for a purring house cat simulates a setting where the event could take place – the event is not simulated in a vacuum. Thus, an interaction with a purring house cat might be simulated in a living room, bedroom, yard, etc. Again such knowledge is represented as simulations, this time as reenactments of particular settings.

In summary, a situated conceptualization typically simulates four basic types of components: (1) perceptions of relevant people and objects, (2) an agent’s actions and other bodily states, (3) introspective states, such as emotions and cognitive operations, and (4) likely settings. Putting all these together, a situated conceptualization is a multimodal simulation of a multicomponent situation, with each modality-specific component simulated in the respective brain area.

It is important to note that a situated conceptualization consists of simulations from many different simulators. A situated conceptualization for a purring house cat is likely to include simulations from simulators for animals, people, objects, actions, introspections, and settings. Thus, a single simulator alone does not produce a situated conceptualization. Instead, many simulators contribute to the collection of components that a situated conceptualization contains.

It is also important to note that situated conceptualizations place the conceptualizer directly in the respective situations, creating the experience of “being there” [Barsalou (2002)]. By reenacting an agent’s actions and introspective states, these complex simulations create the experience of the conceptualizer being in the situation – the situation is not represented as detached and separate from the conceptualizer.

2.3.2. Entrenched situated conceptualizations

Across their life spans, people experience many situations repeatedly in their interactions with people, artifacts, social institutions, etc. As a result, knowledge about these repeated situations becomes entrenched in memory, thereby supporting skilled performance in them. Entrenched knowledge can also guide interactions in novel situations that are similar to these familiar situations [e.g., Andersen and Chen (2002)]. Even
though entrenched knowledge may not always provide a perfect fit, it may often fit well enough to provide useful inferences.

We assume that situated conceptualizations represent people’s entrenched knowledge of these repeated situations. When a situation is experienced repeatedly, multimodal knowledge accrues in the respective simulators for the relevant people, objects, actions, introspections, and settings. The conceptualization’s components become entrenched in the respective simulators, as do associations between these components. Over time, the situated conceptualization becomes so well established that it comes to mind automatically and immediately as a unit when the situation arises. After petting a purring house cat on many occasions, for example, the situated conceptualization for this situation becomes entrenched in memory, such that minimal cuing activates it on subsequent occasions.

2.4. Inference via pattern completion

Once situated conceptualizations become entrenched in memory, they play important roles throughout cognition. In perception, they support the processing of familiar scenes [e.g., Biederman (1981)]. In memory, they support reconstructive retrieval [e.g., Brewer and Treyens (1981)]. In language, they produce situation models and diverse forms of inference [e.g., Zwaan and Radvansky (1998)]. In reasoning, they provide content that facilitates deduction [e.g., Johnson-Laird (1983)]. In social cognition, they provide rich inferences about myriad aspects of interpersonal interaction [e.g., Barsalou et al. (2003)].

2.4.1. Pattern completion with entrenched situated conceptualizations

Much of the processing support that entrenched situated conceptualizations provides appears to result from a pattern–completion inference process. On entering a familiar situation and recognizing it, an entrenched situated conceptualization that represents the situation becomes active. Typically, not all of the situation is perceived initially. A relevant person, setting, or event may be perceived, which then suggests that a particular situation is about to unfold. It is in the agent’s interests to anticipate what will happen next, so that optimal actions can be executed. The agent must draw inferences that go beyond the information given [e.g., Bruner (1957)].

The situated conceptualization that becomes active constitutes a rich source of inference. The conceptualization is essentially a pattern, namely, a complex configuration of multimodal components that represent the situation. When a component of this pattern matched the situation, the larger pattern became active in memory. The remaining pattern components—not yet observed—constitute inferences, that is, educated guesses about what might occur next. Because the remaining components co-occurred frequently with the perceived components in previous situations, inferring the remaining components is justified. When a partially viewed situation activates a situated conceptualization, the conceptualization completes the pattern that the situation suggests. To the extent that a situated conceptualization is entrenched in memory, this process is likely to occur relatively automatically.
Consider the example of seeing a particular cat. Imagine that her face, color, and bodily mannerisms initially match modality-specific simulations in one or more situated conceptualizations that have become entrenched in memory for [CATS]. Once one conceptualization wins the activation process, it provides inferences via pattern completion, such as actions that the cat is likely to take, actions that the perceiver typically takes, mental states that are likely to result, and so forth. The unfolding of such inferences—realized as simulations—produces inferential prediction.

2.4.2. The statistical character of inference

Everything about the production of inferences via pattern completion has a statistical character [e.g., Barsalou (1987, 1989, 1993), Smith and Samuelson (1997)]. Each simulator that contributes to a situated conceptualization is a dynamical system capable of producing infinite simulations [Barsalou (1999b, 2003a,b)]. In a particular situation, each simulation constructed reflects the current state of the simulator, its current inputs, and its past history. An entrenched situated conceptualization is essentially an attractor, namely, an associated collection of simulations that is easy to settle on, because the associations linking them have become strong through frequent use. Infinitely many states near the attractor, however, offer different versions of the same conceptualization, each representing a different adaptation to the situation. Thus, the entrenched conceptualization for interacting with a purring house cat is not just one complex simulation but the ability to produce many related simulations. When encountering the same type of situation on different occasions, the situated conceptualizations that guide an agent vary dynamically, depending on all relevant factors that influence the contributing simulators.

As a consequence, the inferences that arise via pattern completion vary as well. As the conceptualizations that represent a situation vary across occasions, the completions that follow also vary. Somewhat different inferences result from completing somewhat different patterns.

3. Empirical evidence

This section reviews evidence for the two theses of the previous section. On the one hand, accumulating findings implicate modality-specific mechanisms in conceptual processing. On the other, a variety of additional findings suggest that conceptual representations are situated.

3.1. Behavioral evidence for a modal nonmodular conceptual system

Defining the construct of a modality-specific state is useful for assessing whether modal systems underlie conceptual processing. One way to think about such states is as patterns of neural activation. On seeing a cat, its visual representation in the brain includes
patterns of neural activation along the ventral and dorsal streams. Thinking about modality-specific states this way is well established and widely accepted [e.g., Zeki (1993), Palmer (1999)]. Alternatively, modality-specific states can be viewed as conscious mental images. Problematically, though, what becomes conscious is a relatively small subset of the unconscious processing occurring neurally. For this reason, the remainder of this chapter focuses on the neural representation of modality-specific states.

3.1.1. Predictions for modular amodal vs. nonmodular modal theories

Viewing modality-specific states as active neural patterns provides a means of distinguishing these two approaches. According to modular amodal views, a conceptual representation is not a neural pattern in modality-specific systems. The neural patterns that represent an entity during its perception have nothing to do with its conceptual representation. Instead, neural patterns in another brain system represent the object conceptually, using a different format than those in modality-specific systems. Furthermore, amodal representations use the same general format to represent conceptual information about properties from different modalities. Thus, modular amodal views do not predict a priori that modality-specific systems should become active during conceptual processing, nor that different patterns of modality-specific processing should arise for categories having different modality-specific content.

Nonmodular modal views make the opposite predictions. During the conceptual representation of a category, the neural systems that process actual interactions with its instances should also become active as if a category member were present (although not identically). On conceptualizing [CATS], for example, the visual system might become partially active as if a cat were present. Similarly, the auditory system might reenact states associated with hearing a cat, the motor system might reenact states associated with petting a cat, and the limbic system might reenact emotional states associated with enjoying the experience of a cat.

Thus, modular amodal and nonmodular modal theories make different predictions about the roles of modality-specific systems in conceptual processing. Because of these clear differences, it seems that much research in the literature would have addressed which account is correct. Surprisingly, though, little research has addressed this issue. Instead, theoretical considerations have primarily been responsible for the widespread acceptance of modular amodal views throughout the cognitive science community. As Barsalou (1999b) conjectured, the ascendance of the modular amodal approach reflected the development of logic, statistics, and computer science in the twentieth century, and the subsequent incorporation of these developments into the cognitive revolution. Because amodal representation languages have much expressive power, because they can be formalized, and because they can be implemented in computer hardware, they captured the imagination of the cognitive science community, took over theoretical thinking, and became widely practiced.
Nevertheless, a strong empirical case should exist for such a central assumption, even if it is useful theoretically. As the next two subsections illustrate, accumulating findings question this assumption, first from behavioral psychology, and second, from cognitive neuroscience.

3.1.2. Assessing the presence of modality-specific effects in conceptual processing

The behavioral experiments reviewed here used laboratory tasks that are widely assumed to activate and utilize category knowledge. For example, the research in my laboratory has often studied the property generation and property verification tasks. During property generation, a participant hears the word for a category (e.g., *cat*), and then states characteristic properties of the underlying concept out loud (e.g., *claws, whiskers, tail, you scratch it*). During property verification, a participant reads the word for a category on a computer (e.g., *cat*), and then verifies whether a subsequently presented property is true or false for that category (e.g., *claws* vs. *wings*). Whereas property generation is an active, production-oriented task extended over time, property verification is a more passive, recognition-oriented task executed under time pressure.

Most accounts of these tasks assume that participants use amodal representations to perform them [e.g., Kosslyn (1976), Smith (1978)]. When producing or verifying properties, participants access semantic networks, feature lists, frames, etc., to produce the required information. We hypothesized instead that participants simulate a category member to represent a category, and then consult their simulations to produce the requested information. During property generation, participants scan across their simulations, and produce words for properties perceived in them [see Barsalou (2003a), for an account of how properties in the regions of a simulation are perceived]. During property verification, participants evaluate whether test properties can be perceived in these simulations.

Most importantly, these experiments employ the following logic to test the simulation hypothesis: If conceptual processing utilizes modality-specific mechanisms, then modality-specific phenomena should occur during conceptual processing. If variables such as occlusion, size, shape, and orientation affect perceptual processing, they should be likely to affect conceptual processing. Although not all perceptual variables should affect conceptual processing, at least some of them should. In contrast, if participants only use amodal representations during conceptual processing, then it is much less obvious that perceptual variables should affect performance. No amodal theory has ever predicted that variables like occlusion, size, shape, and orientation should affect conceptual processing.

3.1.3. Occlusion during property generation

Wu and Barsalou (2005) offer one example of this experimental logic. In several experiments, they assessed whether occlusion affects conceptual processing. If conceptual processing utilizes perceptual simulations, then a variable like occlusion, which affects vision, might affect conceptual processing as well. Wu and Barsalou manipulated occlusion by asking half the participants to generate properties for noun concepts
(e.g., [LAWN]), and by asking the other half to generate properties for the same nouns preceded by revealing modifiers (e.g., [ROLLED-UP LAWN]). Wu and Barsalou predicted that if people simulate LAWN to generate its properties, they should rarely produce its occluded properties, such as DIRT and ROOTS. As in actual perception, occluded properties should not receive much attention, because they are hidden behind an object’s surface. They also predicted that, conversely, when people produce properties for [ROLLED-UP LAWN], previously occluded properties would become salient in simulations and be produced more often.

Amodal theories of conceptual combination do not readily make this prediction, given that they do not anticipate effects of perceptual variables, such as occlusion. To the contrary, these theories typically assume that conceptual representations abstract over perceptual variation associated with occlusion, size, shape, and orientation. Furthermore, amodal theories of conceptual combination are relatively compositional in nature [e.g., Smith et al. (1988)]. When people combine ROLLED UP with [LAWN], for example, the meaning of the conceptual combination should be roughly the union of the individual meanings. Unless additional post hoc assumptions are added that produce interactions between nouns and modifiers, the properties for [LAWN] are not obviously changed by [ROLLED UP] (e.g., the accessibility of DIRT and ROOTS does not vary).

As Wu and Barsalou predicted, internal properties were produced relatively infrequently for the isolated nouns. Furthermore, the number of internal properties increased significantly when revealing modifiers were present, compared to when they were not present. Internal properties were also produced earlier in the protocols and in larger clusters. This finding occurred both for familiar noun combinations, such as half watermelon, and for novel ones, such as glass car. Rules for properties stored with the modifiers were not responsible for the increase in occluded properties (e.g., perhaps rolled up always increases the salience of unoccluded properties in the head noun). If such rules had been responsible, then a given modifier should have always increased the salience of normally occluded properties. However, in many noun phrases (e.g., rolled-up snake), this was not the case. Occluded properties only increased when the modifiers referred to entities whose internal parts become unoccluded in the process of conceptual combination (e.g., [ROLLED-UP LAWN]). Together, this pattern of results is consistent with the prediction that people construct simulations to represent conceptual combinations. When a simulation reveals occluded properties, they are produced more often, relative to when they remain occluded.

3.1.4. Size during property verification

Solomon and Barsalou (2004) assessed whether perceptual variables such as size affect the property verification task. Participants performed 200 property verification trials, half true and half false. Of primary interest was explaining the variance in response times (RTs) and error rates across the 100 true trials. Why are some concept–property pairs verified faster and with more accuracy than others? To assess this issue, the true
concept–property pairs were scaled for perceptual, linguistic, and expectancy variables that might explain this variance. The linguistic variables included the associative strength between the concept and property words in both directions, the word frequency of the properties, and the word length. The perceptual variables included the size and position of the properties, whether they were occluded, whether they would be handled during situated action, and so forth. The expectancy variables assessed the polysemy of the property words (i.e., property words often have many different senses across objects; consider leg, handle).

The RTs and error rates for the 100 true trials were then analyzed by performing hierarchical linear regression on the three groups of variables. Most notably, the perceptual variables explained significant amounts of unique variance, after variance attributable to the linguistic and expectancy variables had been removed. Within the perceptual variables, the variable of size explained the most unique variance. As properties became larger, they took longer to verify.

This finding suggests that people verify properties by processing the regions of simulations that contain them. As the region that must be processed becomes larger, more time is required to process it. Kan et al. (2003) provided neural corroboration for this conclusion. When participants performed the Solomon and Barsalou experiment in a functional magnetic resonance imaging (fMRI) scanner, activation occurred in the left fusiform gyrus, an area often active in mental imagery and high-level object perception.

3.1.5. Shape during property verification

Solomon and Barsalou (2001) similarly found that detailed property shape was a critical factor in property verification. They assessed whether verifying a property facilitated verifying the same property again later for a different concept. For example, does verifying MANE for [LION] later facilitate verifying MANE for [PONY]? If participants represent MANE with a single amodal symbol that abstracts over differently shaped manes, then verifying MANE for [LION] should activate this symbol so that later, verifying MANE for [PONY] benefits from this. Alternatively, if people simulate manes to verify them, then simulating a lion mane might not later facilitate simulating a pony mane, given their differences in shape. Whereas a lion’s mane wraps around the circumference of its neck, a pony’s mane runs down the length of its neck. Because the shapes of the two manes differ significantly, simulating one might not facilitate simulating the other. Conversely, when the first mane verified has the same detailed shape as the later mane, facilitation should result (e.g., verifying MANE for [HORSE] prior to verifying MANE for [PONY]).

Across several experiments, the results supported the simulation view. When participants verified a property on an earlier trial, it facilitated verifying the same property later, but only if the detailed shape was similar. Thus, verifying MANE for [PONY] was facilitated by previously verifying MANE for [HORSE], but not by previously verifying MANE for [LION].
This effect did not result from greater overall similarity between [HORSE] and [PONY] than between [HORSE] and [LION]. When the property was highly similar for all three concepts, facilitation occurred from both the high and low similarity concepts. For example, verifying BELLY for [PONY] was facilitated as much by verifying BELLY for [LION] as by verifying BELLY for [HORSE]. Thus, the detailed perceptual similarity of the property was the critical factor, not the similarity between concepts. When the detailed shape of the critical property matched across two trials, facilitation occurred. When the detailed shapes differed, they did not. This pattern is again consistent with the conclusion that people simulate properties to verify them.

3.1.6. Modality switching during property verification

Further evidence for this conclusion comes from Pecher, Zeelenberg and Barsalou (2003), who found that a modality-switching phenomenon in perception also occurs during property verification. In actual perception, processing a signal on a modality suffers when the previous signal was perceived on a different modality than when it was perceived on the same modality [e.g., Spence, Nicholls and Driver (2000)]. For example, processing a light flash is faster when the previous signal was a light flash than when it was an auditory tone. A common explanation is that selective attention must shift to a new modality when the modality changes, incurring a temporal cost.

Pecher et al. demonstrated that the same phenomenon occurs during property verification (again using words for concepts and properties). When participants verified a conceptual property on one modality, processing was faster when the previous property came from the same modality than when it came from a different one. For example, verifying LOUD for [BLENDER] was faster when RUSTLING was verified for [LEAVES] on the previous trial than when TART was verified for [CRANBERRIES]. Analogous to perceptual modality switching, this switching cost suggests that people shift between modalities to simulate the properties being verified.

An alternative account is that properties from the same modality have higher associations between them than do properties from different modalities, which produce priming across adjacent trials. When associative strength was assessed, however, properties from the same modality were no more associated than were properties from different modalities. Furthermore, when highly associated properties were verified on contiguous trials in a later experiment, they were verified no faster than unassociated properties. Thus, modality switching and not associative strength appears to be responsible for the obtained effects.

Pecher, Zeelenberg and Barsalou (2004) further demonstrated the modality-switching effect under other task conditions. Marques (in press) extended the conditions that produce this effect, and also obtained this effect in experiments conducted in Portuguese. Barsalou et al. (2005) review the literature on modality switching in property verification. Again, it appears that the process of verifying properties produces simulations of them in modality-specific systems.
3.1.7. Shape and orientation during comprehension

All the evidence reviewed so far for simulation has come from conceptual tasks (i.e., property generation and property verification). Researchers, however, have also found evidence for simulation in language comprehension tasks. Because conceptual representations are widely assumed to underlie the representation of text meaning, these findings further implicate simulation in conceptual processing. The next several sections illustrate some of these recent findings.

Earlier we saw that property shape affects the process of verifying properties [Solomon and Barsalou (2001)]. Zwaan, Stanfield and Yaxley (2002) similarly found that shape affects language comprehension. Participants read a short vignette about an object that implied one of several possible shapes. Some participants read about a flying bird, whereas others read about a sitting bird (i.e., the implied shape of the bird’s wings differed between the two vignettes). After reading the vignette, participants named a picture of an object in isolation, which was sometimes the same as an object just described in the previous sentence. On these trials, the shape of the object was manipulated such that it was either consistent or inconsistent with the implied shape in the sentence. When a bird was shown, for example, sometimes its wings were outstretched, and other times its wings were folded. As the simulation view predicts, participants named objects faster when the pictured shapes matched the implied shapes in the vignettes than when they did not.

Similarly, the implied orientation of an object affects comprehension. In Stanfield and Zwaan (2001), participants read vignettes that implied objects in particular orientations. For example, some participants read about someone pounding a nail into the wall, whereas other participants read about someone pounding a nail into the floor. Immediately afterwards, participants viewed a picture of an isolated object and had to indicate whether it had been mentioned in the vignette. Sometimes the orientation of an object that had occurred in the text matched its implied orientation, and sometimes it did not (e.g., a horizontal nail vs. a vertical one). Verification was fastest when the orientations matched. Again, participants appeared to be simulating objects mentioned in the sentences.

3.1.8. Movement direction in comprehension

In another set of comprehension studies, Glenberg and Kaschak (2002) found that people understood sentences that described actions by simulating the actions in their motor systems. In these experiments, participants read sentences and judged the grammaticality of the sentences. Embedded within the list of sentences were some that described actions moving toward the body (e.g., “Open the drawer.”) vs. others that described actions moving away from the body (e.g., “Close the drawer.”). Participants indicated that a sentence was grammatical by pressing a response button by moving either toward their bodies or away from their bodies. When the button press movements were consistent with the meaning of the sentence, RTs were faster than when they were inconsistent.
For example, participants were faster to verify that “Open the drawer.” is grammatical with a button press toward their bodies than with a button press away from their bodies. Glenberg and Kaschak also observed such effects for sentences that implied an abstract direction of motion. For example, participants were fastest to verify that “Liz told you the story” is grammatical with a button press toward their bodies. This pattern supports the conclusion that participants simulated the meanings of the sentences in their motor systems. When these simulations were consistent with response actions, processing was faster than when they were inconsistent.

3.1.9. Further evidence for simulation from comprehension studies

A variety of other findings further implicate simulation during text comprehension. Glenberg and Robertson (2000) found that readers readily compute the functional affordances of novel objects during comprehension, suggesting that they used perceptual simulations to represent them [also see Kaschak and Glenberg (2000)]. Fincher-Kiefer (2001) asked participants to adopt either a visual or verbal working memory load during text comprehension and found that the visual load produced the greatest interference on a subsequent inference task, suggesting that a simulated situation model represented the text. Richardson et al. (2003) found that the direction of a perceptual stimulus affected the time to process a sentence describing directional motion, suggesting that the sentences’ meanings were being simulated. As these results indicate, accumulating evidence in the comprehension literature implicates simulation in the representation of text meaning.

3.1.10. Behavioral evidence for embodiment in social cognition

Perhaps the largest amount of evidence for simulation comes from social psychology. Because these findings have been reviewed in detail elsewhere, they are simply noted here [see Barsalou et al. (2003), Niedenthal et al. (in press)]. Many studies demonstrate that social stimuli induce bodily states. For example, perceiving various types of people and social events induces postures, arm movements, and facial expressions. The processing of a social stimulus does not simply activate an amodal description of the stimulus in memory. Instead, representations in the modalities become active to play central roles in social meaning.

Many other studies show that bodily states induce high-level, cognitive and emotional social representations. Specifically, various states of the body, arms, head, and face activate social categories, attitudes, and affects. Still other findings demonstrate that social information processing proceeds optimally when cognitive states are compatible with bodily states.

In general, the social psychology literature provides extensive evidence that social cognition is tightly coupled with the modalities, especially the motor, somatosensory, and limbic systems. States of these systems are often implicated in social processing, and have a variety of major effects upon it.
3.2. Neural evidence for a modal nonmodular conceptual system

As the previous section illustrated, a strong behavioral case has been developed for simulations in the conceptual system. A strong case has also been developed in the brain lesion and neuroimaging literatures.

3.2.1. Category-specific deficits

Lesions in a modality-specific system increase the likelihood of losing categories that rely on that system for processing exemplars. Because visual processing is important for interacting with [LIVING THINGS], such as [MAMMALS], damage to visual areas increases the chances of losing knowledge about these categories; because action is important for interacting with [MANIPULABLE OBJECTS], such as [TOOLS], damage to motor areas increases the chances of losing knowledge about these categories [e.g., Warrington and Shallice (1984), Warrington and McCarthy (1987), Damasio and Damasio (1994), Gainotti et al. (1995), Humphreys and Forde (2001)]. Similarly, lesions in color processing areas produce deficits in color knowledge [e.g., DeRenzi and Spinnler (1967)], and lesions in the spatial system produce deficits in location knowledge [e.g., Levine, Warach and Farah (1985)].

This pattern of findings has led many researchers to conclude that knowledge is grounded in the brain’s modality-specific systems. Because the systems used to interact with a category’s members during perception and action produce knowledge deficits when lesioned, category knowledge appears to rely on these systems for representational purposes.

Other factors besides damage to modality-specific systems also contribute to conceptual deficits. Caramazza and Shelton (1998) propose that localized brain areas represent specific categories that are evolutionarily important (e.g., [ANIMALS]). Tyler et al. (2000) propose that the statistical distribution of shared vs. unique property information for categories determines their vulnerability to lesion-based deficits. Thus, theories in this area increasingly include multiple mechanisms for explaining the variety of deficits observed [e.g., Coltheart et al. (1998), Cree and McRae (2003), Simmons and Barsalou (2003)]. Nevertheless, many researchers in this area have concluded that modality-specific systems play central roles in knowledge representation.

3.2.2. Neuroimaging studies of category knowledge

The neuroimaging literature further supports this conclusion [for reviews, see Pulvermüller (1999), Martin (2001)]. Consistent with the lesion literature, different types of categories differentially activate modality-specific systems. Categories that depend heavily on visual information (e.g., [ANIMALS]) strongly activate visual areas during neuroimaging, whereas categories that depend heavily on action (e.g., [TOOLS]) activate the motor system [e.g., Martin et al. (1996)]. Color categories
activate color areas [e.g., Chao and Martin (1999)]. Social categories activate areas central to social interaction [e.g., Decety and Sommerville (2003), Gallese (2003)].

Consider several examples of these studies. Chao and Martin (2000) had participants view briefly presented pictures of manipulable objects, buildings, animals, and faces while lying passively in an fMRI scanner. While participants viewed manipulable objects (e.g., hammers), a brain circuit that underlies the grasping of manipulable objects became active. Notably, this circuit was not active while participants viewed buildings, animals, and faces. In previous studies with both monkey and humans, this grasping circuit became active either when participants actually performed actions with manipulable objects, or when they watched others perform such actions [e.g., Rizzolatti et al. (2002)]. Significantly, this circuit became active even though Chao and Martin’s participants did not move in the scanner, and even though they did not view any agents or actions. Participants simply viewed pictures of static objects in isolation. Because the grasping circuit nevertheless became active, Chao and Martin concluded that this activation constituted a motor inference about how to act on the perceived object. Viewing a manipulable object activated category knowledge about it that included motor inferences (e.g., a hammer can be swung). Most importantly, simulations in the grasping circuit appeared to represent these inferences.

Simmons, Martin and Barsalon (in press) performed an analogous experiment with food categories. While participants lay passively in an fMRI scanner, they viewed food pictures for 2 s, and simply decided whether the current picture was identical to the previous one. Participants were not asked to categorize the foods, nor to think about how they taste. Nevertheless, under these superficial processing conditions, the pictures activated a brain area that represents how foods taste, along with areas that represent the reward value of foods. Even though participants were not actually tasting any foods, these areas became active. As participants perceived a food, it activated category knowledge, which then produced taste inferences via simulations in the gustatory system.

Modality-specific inferences from category knowledge also occur in response to words. Hauk, Johnsrude and Pulvermuller (2004) had participants simply read words for 2.5 s in an fMRI scanner. Within the list, randomly distributed subsets of words referred to head, arm, and leg actions (e.g., lick, pick, and kick, respectively). Hauk et al. predicted that if the meanings of action words are represented as simulations in the motor system, then all three types of words should activate it. Indeed, these words produced activations in motor areas that became active when participants actually moved the respective body parts in the scanner. Most notably, however, the three types of action words differentially activated their respective regions of the motor strip. Words for head actions, arm actions, and leg actions activated the regions that produce head, arm, and leg actions, respectively.

In summary, lesion and neuroimaging results from cognitive neuroscience corroborate the behavioral results reviewed earlier from cognitive and social psychology. These converging bodies of evidence support the first theme of this chapter: The conceptual system utilizes modality-specific mechanisms. Rather than being a modular system that only uses amodal representations, it is a nonmodular system that depends significantly on modal representations in modality-specific systems.
3.3. Evidence for situated conceptualizations

The second theme of this chapter is that conceptual representations are situated. Rather than being a general description of a category, a concept is the productive ability to generate many different situated conceptualizations, each supporting a different course of situated action with the category. As described earlier, a general description of a category used across all occasions would not provide the specialized inferences needed in particular situations. A single representation of [CHAIRS], for example, would be too general to produce the specific inferences needed to interact effectively with dining chairs, office chairs, theater chairs, airplane chairs, or ski lift chairs. Instead, each type of instance is best served by a situated conceptualization tailored to its respective situation.

Barsalou (2003b) proposed that a situated conceptualization supports situated action with a particular category member through four types of situated inferences:

1. Inferences about goal-relevant properties of the focal category;
2. Inferences about the background setting;
3. Inferences about likely actions that the agent could take to achieve an associated goal;
4. Inferences about likely introspective states that the agent might have while interacting with the category, such as evaluations, emotions, goals, and cognitive operations.

To see the importance of these four inferences types, imagine interacting with an airplane chair. Simply activating a general description of [CHAIRS] would provide insufficient inferences about these four aspects of situated action. A general [CHAIR] concept would not predict: (1) the particular parts of an airplane chair, (2) relevant aspects of the setting, (3) actions that could be performed on an airplane chair, and (4) introspections that might result.

Alternatively, activating a situated conceptualization supports all four types of prediction, including: (1) an airplane chair has controls for adjusting the seat back angle and headphone volume; (2) an airplane chair resides in a crowded setting with little space between adjacent chairs, so that tilting back one’s chair impinges on the space of the passenger behind; (3) the action of pressing a light button on the chair activates an overhead reading light; (4) sitting in an airplane chair produces introspections that include negative affect about being in a cramped setting. As described earlier, these inferences are delivered via a pattern completion process that operates on situated conceptualizations. When a familiar category member is categorized, a situated conceptualization containing it becomes active and produces the four types of situated inferences.

The next four subsections review evidence that the conceptual system delivers these four types of situated inferences as people activate concepts. Rather than generic concepts becoming active, sets of situated inferences become active instead. Each of the next four subsections provides evidence in turn for one type of situated inference: goal-relevant properties of the focal category, background settings, actions, and introspective states.
3.3.1. Inferences about goal-relevant properties of the focal category

Many studies demonstrate that concepts do not produce the same generic representation over and over again across situations. Instead, a concept produces one of many possible representations tailored to the current context.

Barsalou (1982) illustrates this general finding. After participants had read a sentence, they verified whether a subsequent property was true or false of the subject noun. As the following examples illustrate, the predicate of the sentence varied between participants to manipulate the context of situated action:

The basketball was used when the boat sank.
The basketball was well worn from much use.

As can be seen, the predicate in each sentence situates [BASKETBALL] in a different context. Immediately after reading one of these two sentences, participants verified whether FLOATS was a true property of [BASKETBALL]. As the situated conceptualization view predicts, participants verified FLOATS 145 ms faster after reading the first sentence than after reading the second. As participants read about the boat sinking, the concept for [BASKETBALL] produced relevant inferences, such as FLOATS. The concept for [BASKETBALL] did not produce the same representation in both contexts.

Many additional studies across multiple literatures have found similar results [for a review, see Yeh and Barsalou (in press)]. In memory, context effects on word encoding are widespread [e.g., Greenspan (1986)]. In category learning, background knowledge about a situation constrains the properties of objects salient for them [e.g., Murphy (2000)]. In sentence processing, context effects on lexical access are legion [e.g., Kellas et al. (1991)]. Such findings clearly demonstrate that a general description of a category does not represent the category across situations. Instead, its representation is tailored to current task conditions.

3.3.2. Evidence for setting inferences

When the conceptual system represents a category, it does not do so in a vacuum. Instead, it situates the category in a background setting. As Yeh and Barsalou (in press) review, much work supports the inclusion of setting inferences in category representations.

Vallée-Tourangeau, Anthony and Austin (1998) provided one example of this evidence. On each trial, participants received the name of a category (e.g., [FRUIT]) and produced instances of it (e.g., [APPLE], [KIWI], [PEAR]). Accounts of this task typically assume that participants generate instances from conceptual taxonomies, or from similarity-based clusters. For [FRUIT], participants might first produce instances from [CITRUS FRUIT], then from [TROPICAL FRUIT], and then from [WINTER FRUIT]. Alternatively, Vallée-Tourangeau et al. proposed that participants situate the category in a background setting, scan across the setting, and report the instances present. To produce [FRUIT], for example, participants might imagine being in the produce section of their grocery store and report the instances found while scanning through it. Notably,
this prediction assumes that categories are not represented in isolation. Instead, categories are associated with situations, such that categories and situations become active together in situated conceptualizations.

To assess this hypothesis, Vallée-Tourangeau et al. first asked participants to produce the instances of common taxonomic categories (e.g., [FRUIT]) and also of ad hoc categories (e.g., [THINGS PEOPLE TAKE TO A WEDDING]). After producing instances for all of these categories, participants were asked about the production strategies they had used, indicating one of three possible strategies for each category. First, if a category’s instances came to mind automatically, participants indicated the unmediated strategy. Second, if the instances were accessed according to clusters in a taxonomy, participants indicated the semantic strategy. Third, if the instances were retrieved from experienced situations, participants indicated the experiential strategy.

As Vallée-Tourangeau et al. predicted, retrieving instances from situations was the dominant mode of production. Participants indicated they used the experiential strategy 54% of the time, followed by the semantic strategy (29%) and the unmediated strategy (17%). Surprisingly, this pattern occurred for both taxonomic and ad hoc categories. Because ad hoc categories arise in goal-directed situations, it is not surprising that they would be situated. Surprisingly, though, taxonomic categories were situated as well. Walker and Kintsch (1985) and Bucks (1998) reported similar findings.

The Wu and Barsalou (2005) experiments on occlusion described earlier further demonstrate that categories are situated in background settings. In those experiments, participants were explicitly instructed to produce properties of the target objects (e.g., [WATERMELON]). The instructions neither requested nor implied the relevance of background settings. Nevertheless, participants produced setting information regularly. Across experiments, the percentage of setting information ranged from 19% to 35%, averaging 25%. As participants simulated the target objects so that they could generate properties, they implicitly situated the objects in background settings, leading to the inadvertent production of many setting properties (e.g., PARK, PICNIC TABLE, etc., for [WATERMELON]).

Many further findings demonstrate a tight coupling between object representations and settings [as Yeh and Barsalou (in press) review in greater detail]. For example, studies of visual object processing have frequently shown that objects are strongly associated with their background scenes [e.g., Biederman (1981)]. On perceiving a familiar isolated object, a typical background scene is inferred immediately.

3.3.3. Evidence for action inferences

The previous two subsections illustrated that conceptual representations contain contextually relevant properties, and that they are situated in background settings. As next subsection illustrates, these situated representations of categories are not represented as detached from the conceptualizer. Instead, the conceptual system places the conceptualizer in these situated representations, producing inferences about possible actions the conceptualizer could take. In other words, the conceptual system implements simulations of “being there” with category members [Barsalou (2002)].
We have already seen a number of findings that support the presence of such inferences. As Glenberg and Kaschak (2002) demonstrated, reading a sentence about an action (with no agent mentioned) activates a motor representation of it. Rather than representing the action in a detached amodal manner, the action is represented as if the conceptualizer were preparing for situated action. The embodiment effects from social psychology mentioned earlier are also consistent with this conclusion. Relevant neuroimaging evidence was also reviewed. In Chao and Martin (2000), the grasping circuit became active when participants viewed manipulable artifacts in isolation. In Hauk et al. (2004), simply reading an action word activated the relevant part of the motor strip. All of these findings are consistent with the proposal that the conceptual system is action oriented – it is not simply a repository of amodal descriptions. Thus, when a concept becomes active, it prepares the conceptualizer for interacting with its instances by priming relevant actions in the motor system.

Adolphs et al. (2000) provide further evidence for this claim. On each trial, participants viewed a picture of face and indicated whether the expression was happy, sad, angry, etc. To assess the brain areas responsible for these categorizations, Adolphs et al. sampled from a registry of patients having lesions in different brain areas. To the extent that an area is important for categorizing visual expressions of emotion, lesions in the area should produce task deficits. The important finding was that large deficits resulted from lesions in the somatosensory cortex. Why would somatosensory lesions produce deficits on a visual task? Adolphs et al. proposed that simulating facial expressions on one’s own face is central to visually recognizing expressions on other faces. When the somatosensory cortex is damaged, these simulations become difficult, and facial categorization suffers.

Work in social cognition corroborates this conclusion. Wallbott (1991) videotaped participants’ own faces as they categorized emotional expressions on other’s faces. Notably, participants simulated the emotional expressions that they were categorizing on their own faces. Furthermore, participants’ accuracy was correlated with the extent to which their facial simulations were recognizable. Niedenthal et al. (2001) similarly found that preventing participants from simulating facial expressions decreased their ability to categorize other’s facial expressions. Together, these findings illustrate that action systems in the brain become involved in the visual processing of faces.

Facial simulations of emotional expression in others can be viewed as motor inferences that the conceptual system produces to support situated action. Under many conditions, if another person is experiencing an emotion, it is often useful for the perceiver to adopt the same emotional state. Thus, if another person is happy about something, it is often supportive to be happy as well. Once the concept for a particular emotion becomes active, appropriate motor and somatosensory states for adopting it oneself follow as conceptual inferences.

3.3.4. Evidence for introspective state inferences

As we just saw, the conceptual system inserts the conceptualizer into situated conceptualizations via the simulation of possible actions in the situation. As this next
subsection illustrates, the conceptual system further inserts the conceptualizer into situated conceptualizations via the simulation of possible introspections. As a particular situation associated with a category is simulated, relevant emotions, evaluations, goals, and cognitive operations are included.

Again, consider findings from the Wu and Barsalou (2005) occlusion experiments. Earlier we saw that participants produced properties about background settings, even though the explicit instruction was to produce properties of the target objects (e.g., \[\text{WATERMELON}\]). Of interest here is the fact that participants also produced many properties about the likely introspective states that they would experience in associated situations. These included evaluations of whether objects are good, bad, effective, ineffective, etc. They also included emotional reactions to objects, such as happiness, along with other cognitive operations relevant to interacting with them (e.g., comparing an object to alternatives). On the average, 10% of the properties that participants produced were about introspective states, ranging from 6% to 15% across experiments. As participants simulated a target object, they situated it in a background setting and included themselves as agents. In the process, they simulated likely introspective states that they would experience, which they then inadvertently described in their protocols.

Adopting particular perspectives in laboratory tasks further implicates introspective states in conceptual simulation. A perspective can be viewed as an introspective state because it reflects one of many possible views that an agent could take in a physical situation. As people simulate situations, they imagine likely perspectives. In Spivey et al. (2000), participants listened to a vignette while wearing an eye-tracking helmet (which they believed was turned off at the time). Whereas some participants heard about a skyscraper, others heard about a canyon. As participants listened to the vignette, they tended to adopt the relevant perceptual perspective on the setting. Participants hearing about the skyscraper were most likely to look up, whereas participants hearing about the canyon were most likely to look down. Participants acted as if they were “there,” adopting the relevant perspective on the situation.

In a related study, Barsalou and Barbey (2005) videotaped participants as they produced properties for object concepts. On a few trials, the object was something that would typically be encountered above a person (e.g., \[\text{BIRD}\]) or on the ground (e.g., \[\text{WORM}\]). When participants produced properties for objects typically found above them, their eyes, face, and hands were more likely to drift up than for objects typically found below them, and vice versa. Again, participants appeared to simulate the perspective on the situation that they would take if interacting with the object.

3.3.5. Evidence for dynamical simulations

As we have seen, a general decontextualized description does not represent a category across situations. Instead, conceptual representations appear to be highly contextualized, containing various types of specialized inferences that support situated action in specific contexts. If this account is correct, then another prediction follows: Many different representations of a category should be observed both between different individuals and in
one individual. When different individuals represent a given category, their representations of it should differ, depending on the situation that they are anticipating. Similarly, when the same individual represents a category on different occasions, its representations should again differ, depending on the anticipated situation.

A variety of findings support this prediction [as reviewed in Barsalou (1987, 1989, 1993)]. Consider the agreement for typicality judgments across different members of a category (e.g., the typicality of different birds). Across a variety of studies, the average correlation in typicality judgments between pairs of participants for a given category was only around 0.40. Different participants appeared to use very different prototypes for judging typicality. Additionally, individual participants appeared to use different category representations on different occasions. When the same participant judged typicality 2 weeks after an initial judgment, the average correlation with their earlier judgment was about 0.80, indicating a change in how they represented the categories.

Similar variability arises in other tasks. In McCloskey and Glucksberg (1978), participants assigned basic-level categories to superordinates in two sessions separated by a month. Across the two sessions, roughly 25% of the basic-level categories changed superordinate membership, indicating variability in categorization criteria. In Barsalou (1989), participants exhibited variability in property generation. On average, two participants only produced 44% of the same properties for the same category. Over a two-week delay, the same participant only produced 66% of the same properties.

On the basis of these results, Barsalou (1987, 1989, 1993) concluded that a concept is a dynamic system. Depending on a variety of conditions, a concept produces a wide variety of different representations. After reviewing similar findings from the conceptual development literature, Smith and Samuelson (1997) reached a similar conclusion. Together, these results challenge the view that a concept is a general description used over and over again across situations. Instead, a concept appears to be an ability or skill to construct specific representations that support different courses of situated action. Because a concept produces a wide variety of situated conceptualizations, substantial variability in its representation arises.

4. Conclusion

Increasing empirical evidence supports the two themes of this chapter. First, the conceptual system does not appear to be modular, nor to solely use amodal representations. Instead, it appears to share mechanisms with modality-specific systems, such that its representations are often modal. Second, the conceptual system does not traffic solely in general descriptions of categories that remain stable across contexts. Instead, it dynamically produces contextualized representations that support situated action in different situations.

The construct of a situated conceptualization integrates these two themes, where a situated conceptualization is a multimodal simulation that supports one specific course of goal-directed interaction with a particular category instance. A given concept
produces many different situated conceptualizations, each tailored to a different instance and setting. A situated conceptualization creates the experience of “being there” with a category instance via integrated simulations of agents, objects, settings, actions, and introspections. On recognizing a familiar type of category instance, an entrenched situated conceptualization associated with it becomes active to provide relevant inferences via pattern completion.

4.1. Important issues for future research

Clearly, many outstanding issues remain, including the role of amodal symbols in the conceptual system, the implementation of classic symbolic functions, and the representation of abstract concepts. Each is addressed briefly in turn.

4.1.1. Amodal symbols

One obvious question is whether amodal symbols coexist with simulations in the conceptual system. Perhaps amodal symbols are necessary to perform classic symbolic functions that simulations cannot implement. Barsalou (1999b, 2003a), however, argues that a simulation-based system can, at least in principle, implement these functions. It remains an open empirical question how these functions are implemented, although some evidence suggests that simulations contribute to them (as discussed shortly).

One possibility is that the classic amodal symbols found in predicate calculus-based approaches to representation do not exist in the conceptual system, but that other types of (relatively) amodal symbols do. For example, the sets of conjunctive neurons in association areas that trigger simulations in feature systems could be viewed as amodal vectors [Simmons and Barsalou (2003)]. Damasio (1989), however, argues that conjunctive neurons serve only to control simulations – they do not function as stand-alone representations. Simmons and Barsalou (2003) further propose that conjunctive neurons have modality-specific tunings, such that they are not truly amodal.

Another obvious issue is whether all conceptual representations are situated. Although the focus here has been on situated conceptualizations, abstractions appear to occur ubiquitously throughout conceptual processing. During everyday cognition, people appear to frequently process generalizations about categories, such as “Hondas are reliable,” “Reality TV shows are boring,” and “Academics are liberal.” Barsalou (2003a) argues that such abstractions are central to human cognition, and that it is essential for simulation theories to explain them. In this spirit, Barsalou (2003a) proposes that simulators implement abstractions via productively constructed configurations of simulations. Much future work is clearly needed to determine whether complex simulations do indeed underlie abstractions. It is also important to establish how abstractions and situated conceptualizations interact during conceptual processing.
4.1.2. Symbolic functions

Another important issue concerns whether simulations implement classic symbolic functions, such as the type-token distinction, categorical inference, productivity, and propositions. Barsalou (1999b, 2003a) argues that, at least in principle, simulations can implement these functions. Preliminary empirical evidence suggests that simulations are involved to some extent. For example, evidence reviewed earlier suggests that the process of predicating a property of a category—a central symbolic function—relies on simulations. When people perform the property verification task, perceptual variables such as size and shape affect performance [Solomon and Barsalou (2001, 2004)]. The time to verify a property becomes shorter the smaller a property is and the more its shape matches property simulations run earlier. If people verify properties by processing regions of simulations, size and shape would have these sorts of effects.

We also saw evidence earlier that simulations underlie conceptual combination, another central symbolic function. Specifically, Wu and Barsalou (2005) found that occlusion affected the properties produced for conceptual combinations. When properties were occluded in isolated nouns, they were produced infrequently. When properties were unoccluded in conceptual combinations, their production increased. This finding suggests that simulations are used to combine modifier and noun concepts.

Clearly, much more work is necessary to establish how the brain implements classic symbolic functions. Preliminary evidence, though, suggests that simulations play a role.

4.1.3. Abstract concepts

Many people find it intuitively easier to understand how the simulation view explains concrete concepts than how it explains abstract concepts. Barsalou (1999b), however, proposed that the conceptual content of abstract concepts is drawn from events and introspections, and that this content could be simulated in relevant modality-specific systems. Rather than being redescribed with amodal symbols, the content of abstract concepts could be simulated just as for concrete concepts, with the difference being in the content simulated.

Barsalou and Wiemer-Hastings (2005) provided preliminary evidence for this view. When participants generated properties for abstract and concrete concepts, they tended to produce situated content for both, including agents, objects, settings, actions, and introspections. Indeed, the distributions of content were remarkably similar. As predicted, the difference was that the concrete concepts contained more content about physical objects, background settings, and simple behaviors, whereas the abstract concepts contained more content about people, social interactions, complex relations, and introspections. In general, the content reported for abstract concepts was drawn from the four types of content for situated conceptualizations reviewed earlier. Thus, it appears possible that this content could be simulated.

Clearly, much more work on abstract concepts is needed to understand how the brain represents this fundamentally important type of concept. Nevertheless, simulation may play a central role.
Acknowledgment

This work was supported by National Science Foundation Grants SBR–9421326, SBR–9796200, SBR–9905024, and BCS–0212134 to Lawrence W. Barsalou.

References


Chapter 29

PERCEPTUAL AND SEMANTIC REORGANIZATION DURING CATEGORY LEARNING

ROBERT L. GOLDSTONE
BRIAN J. ROGOSKY
RACHEL PEVTZOW
MARK BLAIR
Indiana University

Contents

Abstract 652
1. Introduction 653
2. Concept learning and perception 653
   2.1. Object segmentation 654
   2.2. Experiment 1
      2.2.1. Method 655
      2.2.2. Results and discussion 657
   2.3. Experiment 2
      2.3.1. Method 659
      2.3.2. Results and discussion 660
   2.4. Conclusions on perceptual reorganization 662
3. Semantic reorganization during category learning 664
   3.1. Integral versus separable dimensions 664
   3.2. Experiment 3
      3.2.1. Method 665
      3.2.2. Results and discussion 667
   3.3. Experiment 4
      3.3.1. Method 669
      3.3.2. Results and discussion 669
4. Conclusions on semantic reorganization 671
5. Integrating perceptual and semantic reorganization 671
   5.1. Characterizing psychological features 672
   5.2. Characterizing featural change 673
   5.3. Prospects for synthesizing perceptual and semantic reorganization 675
Acknowledgments 676
References 676

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved
Abstract

Category learning not only depends upon perceptual and semantic representations; it also leads to the generation of these representations. We describe two series of experiments that demonstrate how categorization experience alters, rather than simply uses, descriptions of objects. In the first series, participants first learned to categorize objects on the basis of particular sets of line segments. Subsequently, they were given a perceptual part-whole judgment task. Categorization training influenced participants’ part-whole judgments, indicating that whole objects were more likely to be broken down into parts that were relevant during categorization. In the second series, correlations were created or broken between semantic features of word concepts (e.g., ferocious vs. timid, and group-oriented vs. solitary animals). The best transfer was found between category learning tasks that shared the same semantic organization of concepts. Together, the experiments support models of category learning that simultaneously create the elements of categorized objects’ descriptions and associate those elements with categories.
1. Introduction

Human concept learning clearly depends upon the descriptions we give to the objects we categorize. Our concept of [DOG] is built out of features such as “furry,” “barks,” “four-legged,” “domesticated,” “friendly,” and “loyal.” However, recent research has found that the dependency works both ways. People’s object representations not only influence, but are influenced by, the concepts that they learn. We have been exploring the psychological mechanisms by which concepts and descriptions mutually influence one another, and building computational models to show that the circle of influences is benign rather than vicious. Our efforts are not solitary. There is a growing body of behavioral [Shiffrin and Lightfoot (1997), Gauthier et al. (1998), Livingston, Andrews and Harnad (1998)], developmental [Needham (1999)], neural [Kaas (1991), Gauthier and Tarr (1997), Sigala, Gabbiani and Logothetis (2002) Gauthier et al. (2003)] and computational [Rumelhart and Zipser (1985), Hofstadter and Mitchell (1994), Harnad, Hanson and Lubin (1995), Behrmann, Zemel and Mozer (1998), Palmeri, Wong and Gauthier (2004)] evidence suggesting that it is necessary and desirable to develop categories and descriptions for objects simultaneously.

In the [DOG] example above, we purposefully merged what might be thought to be two different kinds of descriptions – perceptual and semantic. We aim to develop a unified account of perceptual and semantic reorganization that accompanies category learning. This is consistent with our larger effort to reunite perceptual and conceptual processes [Goldstone (1994), Goldstone and Barsalou (1998)]. In what follows, we describe two series of experiments implicating category learning in representational reorganization. The first series focuses on a case of perceptual reorganization, while the second focuses on semantic reorganization. However, similar mechanisms are likely to underlie both kinds of reorganization, encouraging the effort to unite perceptual and conceptual adaptation processes.

2. Concept learning and perception

Within traditional work on concept learning and categorization, there has been little suggestion that learned concepts influence perception. A working assumption made by many of the most influential theories of categorization [Bruner, Goodnow and Austin (1956), Medin and Schaffer (1978), Hintzman (1986)] is that objects to be categorized are described along a fixed set of features. The categorization procedure uses, but does not alter, the perceptual descriptions.

However, recently a number of researchers have argued that in many situations, the categorization process influences the featural descriptions that are used. Rather than viewing the “vocabulary” of primitives as fixed by low-level processes, this view maintains that the vocabulary is dependent on the higher-level processes that use the vocabulary. Some evidence for this comes from the study of expert/novice differences. Evidence suggests that experts perceive structures in X-rays [Lesgold et al. (1988), Norman et al. (1992), Sowden, Davies and Roling (2000)], beers [Peron and Allen (1988)], and infant chickens
[Biederman and Shiffrar (1987)] that are missed by novices. Experts in these fields seem to acquire new ways of perceptually structuring objects as they learn new concepts.

2.1. Object segmentation

Objects often have more than one possible segmentation. The letter X can be viewed as composed of two crossing diagonal lines, or as a V and an upside-down V that just touch at their vertices. Segmenting objects into parts is an important part of the process of object recognition [Hoffman and Richards (1984), Hummel and Biederman (1992)]. Stephen Palmer (1977) argued that some segmentations of an object into parts are psychologically more natural than others. He developed a set of measures for determining the naturalness of a particular segmentation of an object. In one measure, Palmer assumed that the longer it took participants to verify whether a particular part was contained in an object, the less natural was the part. For example, in Figure 1, participants saw the whole object on the left and one of the four parts on the right. Participants would generally take longer to respond that the unnatural parts belonged to the whole than that the natural parts did. In general, Palmer’s different measures of segmentation naturalness closely converged. Parts that were natural according to one measure were usually found to be natural according to other measures as well. Furthermore, the measures agreed well with a formal model of part naturalness that integrates several different sources of physical information. In this model, natural object parts tend to have components that are close to each other, have similar orientations, are connected to each other, and have similar lengths.

Our experiments used materials and tasks similar to those used by Palmer, and examined the possibility that information that is physically present in an object is not sufficient to determine its segmentation into parts. Rather, information about a person’s categorization experience may also be necessary to determine the most natural segmentation.

![Figure 1](image_url)

**Fig. 1.** The whole on the left can be segmented into either natural parts or unnatural parts.
2.2. Experiment 1

Experiment 1 tests whether categorization training can alter the naturalness of a part within a whole, as measured by part-whole response times. Participants’ categorization experience is manipulated by giving them one of two different categories to learn. Both groups of participants are then given the same set of part-whole judgments.

The categorization conditions differ in the set of line segments that are diagnostic for categorization. The stimuli to be categorized are distorted versions of Objects A, B, C, and D in Figure 2. For one group of participants, A and B are placed in one category, and C and D are placed in another category. For this group of participants, the three line segments that comprise Part E and the three line segments that comprise Part F are diagnostic for categorization. Objects that belong in one category all have Part E, and objects that belong to the other category all have Part F. For the second group of participants, A and C are placed in one category, and B and D are placed in another category. For these participants, Parts G and H are diagnostic for categorization.

Categorization training could influence later part-whole judgments by highlighting segmentations of whole objects that involve diagnostic parts. For example, if Part F in Figure 2 was diagnostic during categorization training, then participants may be able to

![Fig. 2. Materials used in the categorization portion of Experiment 1. The four objects A, B, C, and D are categorized into two groups. Four other objects (not shown) are categorized into a third “junk” group. When A and B are placed in one group, and C and D are placed in the other, Parts E and F are diagnostic for the categorization. When A and C are placed in one group, and B and D are placed in the other, then Parts G and H are diagnostic.](image-url)
decide relatively quickly that Part F is contained in the whole object in Figure 1, even though it would be considered by Palmer’s quantitative model of part goodness to be relatively unnatural. Experiment 1 tests for the influence of categorization training by comparing the part-whole judgments involving Parts E and F in Figure 2 to those involving G and H, as a function of the categorization training condition.

In Experiment 1, category parts and complements of those category parts are tested. A category part is defined as one of the sets of three line segments that were used to construct the four objects to be categorized in Figure 2. Parts E, F, G, and H are all category parts. Category parts can either be diagnostic (if they are relevant for the categorization) or nondiagnostic. The complement of a part is defined as the line segments that remain after the category parts are removed from a whole. Figure 3 shows the four possible types of trials. On “Present Category Probe” trials, participants are probed with a category part that is present in the whole. On “Absent Category Probe” trials, participants are probed with a category part that is not present in the whole. On “Present Complement” trials participants are probed with a complement (all of the line segments except those belonging to the category part) that is present. On “Absent Complement” trials, a randomly chosen complement to a category part within another whole is used as a probe.

2.2.1. Method

There were two tasks in the experiment: categorization and whole-part decisions. In the categorization phase of the experiment, 49 participants were shown distortions of Objects A, B, C, and D as shown in Figure 2. Distortions of these objects were created by adding one line segment at a random location so that it was connected to at least one other line. Participants were asked to categorize an object into one of three groups. Following the response, a check mark was displayed if the participant was correct, or
an X appeared if the participant was incorrect. In one categorization condition, Objects A and B belonged to one category and Objects C and D belonged to the other category. In the other categorization condition, Objects A and C belonged to one category, and Objects B and D belonged to the other category.

In the second phase of the experiment, trials consisted of displays with “wholes” and “probes.” The participants’ task was to decide whether the probed part was contained in the whole. The wholes consisted of one of the four category-defining parts (E, F, G, or H) from the categorization task, plus three connected line segments (complements), which were connected to the category part. The complements had no lines overlapping any of the category parts. The probes were either category parts (nondiagnostic or diagnostic) or complements.

There were four types of trials in the whole-part decomposition task: present category probe, absent category probe, present complement, and absent complement. For each of the trials shown in Figure 3, the object on the left is the whole, and the object on the right is the probe. In the first type of trial, the probe is a category part that is contained within the whole object. In the second, the probe is the complement to the category part. In the absent category probe trials, the probe is a category part, but is not contained within the whole object. For the last type of trial, absent complement, the probe is a randomly chosen complement from another object. Wholes were presented alone for 1000 ms, and then a probe was added to the display. The participants’ task was to decide, as quickly and accurately as possible, whether or not the whole contained the part.

2.2.2. Results and discussion

Figure 4 shows the mean response times to decide whether or not the part was present in the whole, as a function of whether or not the whole contained a diagnostic category part. Response times to respond to category parts were faster for wholes that contained a diagnostic category part than for those that contained a nondiagnostic part. This diagnosticity advantage was significant only for present category parts. For complements, responses were faster for present than absent complements.

The results indicate that category learning influences perceptual sensitivity. Participants were more sensitive at responding to parts within whole objects when those parts were diagnostic. “Present” response times were significantly lower for diagnostic than nondiagnostic parts, and “absent” response times tended (nonsignificantly) to be lower as well.

2.3. Experiment 2

Experiment 2 further explores the hypothesis that category learning alters the subsequent segmentation of objects into parts. Experiment 2 introduces a new control for category parts: mirror-image reflections of category parts. During the part-whole judgment task, participants were presented with category parts in some trials and reflections of category parts in other trials. Figure 5 shows six types of trials that were used.
On “Present Category Part” trials, participants were presented with wholes that contained parts that were either diagnostic or nondiagnostic during categorization. On “Present Reflection of Category Part” trials, participants were presented with wholes and parts that were horizontal reflections (mirror images) of the category part trials. Finally, other parts were also tested that were neither category parts nor reflections of category parts.

Reflections of category parts are useful controls because the naturalness of a part within a whole remains invariant under reflection in Palmer’s (1977) model of part goodness. For example, whatever the naturalness of Part P is in Whole W, Palmer’s model predicts that the reflection of P should have the same naturalness in the reflection of W. Palmer’s features for naturalness (e.g., cohesion, similarity, and proximity of line segments) remain unchanged if both the whole and the part are rotated or reflected in the same manner. In Figure 5, the “Present Category Probe” and “Present Reflection of Category Probe” conditions are predicted by Palmer’s model to be equally difficult.

However, if category learning can alter the way in which an object is segmented, then it should be possible to change the quality of a part within a whole without much change to the quality of the part’s reflection within the whole’s reflection. If the top part of Figure 5 is diagnostic for categorization, then participants may be able to decide relatively quickly that the top whole contains this part.

Fig. 4. Results of Experiment 1. Line segments were more readily identified as present in whole objects when they were diagnostic during categorization training than when they were nondiagnostic.
2.3.1. Method

The procedure for Experiment 2 was similar to that of Experiment 1. Fifty-seven participants under the experimental conditions were given categorization training followed by a part-whole judgment task. The categorization training was identical to that of the first experiment, using the same stimuli (Figure 2). The two experimental conditions were identical to the two groups in the first experiment. A third group of 38 participants served as controls and received no categorization training.

The part-whole judgment task differed only slightly from that in Experiment 1. A new condition was added, in which the whole and the part were reflected. In addition, the parsing of an object was different: the probe was either a category part or a “cross-pars” part. Figure 5 shows examples of the different types of probes and trials. In the top two examples in Figure 5, the probe is a category part. These types of trials are identical to their comparable trials in Experiment 1. The middle two trials are similar to the top trials, in that the probes are category parts. However, unlike the top trials, the whole and the probe have been reflected (i.e., flipped horizontally). The last two examples of
trials are present and absent cross-parse probes. For “present cross-parse” trials, the parsing of the whole into the cross-parse part is incompatible with the parsing required for “present category part” trials. The cross-parse cuts across the parsing needed to identify the category part, because the cross-parse part has an overlapping line segment in common with the category part. When a cross-parse probe is present, it shares a line with the category part contained within the whole. Absent cross-parse probes do not share any lines with the category part contained within the whole object; rather, they share a common line with one of the category parts that is not present within the whole. While complement parts (Experiment 1) were the remains of the whole after a category part was removed, the cross-parse parts used in Experiment 2 shared one line in common with the category part.

In this experiment, there were five factors of interest: type of probe (category or cross-parse), diagnosticity of the category part contained within the whole (diagnostic or nondiagnostic), diagnosticity of probe (diagnostic or nondiagnostic), trial type (present or absent), and reflection (normal or reflected stimuli). The two values along each of the factors occurred with equal frequency.

2.3.2. Results and discussion

Figure 6 shows the mean response times to decide whether or not the part was present in the whole, as a function of whether or not the whole contained a diagnostic category part. The baseline response times obtained from the control (no categorization) participants for the different types of probes were subtracted from the other conditions. By subtracting out this baseline, we control for differences between the category parts and cross-parse parts in terms of intrinsic naturalness. The response times in Figure 6 are negative because the control group generally took longer to respond than the categorization groups. Thus, lower negative numbers are associated with greater advantages over the control group. Considering only trials in which the probe was a normal category part, participants were faster to respond “present” when the whole contained a diagnostic category part than when it contained a nondiagnostic part. This result replicates Experiment 1, in which participants were faster to respond to diagnostic than nondiagnostic category probes.

Diagnosticity had a significant effect on present, normal cross-parse probes. For this type of probe, response times were slower when the whole object contained a diagnostic category part than when it contained a nondiagnostic part. This is the opposite of the effect found for category probes, for which response times decreased for wholes containing diagnostic compared to nondiagnostic category parts.

Reflecting the stimuli had an effect on present, nondiagnostic category parts and cross-parse probes. When the whole contained a nondiagnostic category part, times to respond “present” were slower for normal category probes than for reflected ones. For both diagnostic and nondiagnostic absent cross-parse probes, response times were lower when the stimuli were reversed than when they were normal.

Categorization training had reliable effects on subsequent part-whole judgments, consistent with the position that participants tend to segment objects into parts that have
been useful during categorization. The most straightforward effect of diagnosticity is on trials where a normal (not reflected) category part is present in the whole object, and participants are probed with this category part. On these trials, if the part was diagnostic during categorization, participants are faster to respond than if it was nondiagnostic.

The positive influence of diagnosticity of category parts was not found for horizontal reflections (mirror images) of the category parts. This result indicates a lack of transfer from learning about one part to other similar parts. A part and its reflection share commonly posited emergent features such as closure, angularity, length, height/width ratio, and density. The lack of transfer to reflected parts suggests that categorization learning sensitizes the particular three-line segments that are diagnostic rather than general stimulus properties of the diagnostic parts.

Fig. 6. Results from Experiment 2. Category parts were more quickly identified as present in whole objects when they were diagnostic during categorization training. Conversely, cross-parse parts were more quickly identified as present in whole objects when the whole objects contained a part that was nondiagnostic during categorization. Asterisks denote significant effects of diagnosticity.
The second influence of categorization training was that, if a whole object contained a diagnostic part, then responses to present noncategory parts were slowed. In other words, on some trials, a whole object contained both a part that was relevant during categorization training and an additional part that had never been seen during categorization. If participants were probed with the never-before-seen part, they were relatively slow to respond “present.” The critical aspect of the stimulus design that may explain this result is that category parts and cross-parse parts always shared one line segment. For example, in Figure 5, the category part in the top panel and the cross-parse part in the bottom panel have one line segment in common. Consequently, any segmentation that involved the category part was incompatible with the segmentation that involved the cross-parse part. If category learning biased participants to see the whole as containing the category part, then we would expect other, inconsistent segmentations of the object to be inhibited. Even though diagnosticity has a harmful influence on part-whole judgments involving new parts, this effect is consistent with the positive influence of diagnosticity. In short, object segmentations that are consistent with previously learned parts are facilitated, and those that are inconsistent with previously learned parts are inhibited.

2.4. Conclusions on perceptual reorganization

These two experiments are generally consistent in indicating that concept learning influences later perceptual part-whole judgments. Participants were more quickly able to identify parts as being present in an object when they were relevant, rather than irrelevant, for an earlier categorization task. This effect could not be explained by a bias to respond “present” because “absent” responses were never slowed, and were sometimes facilitated, for diagnostic parts. The pattern of results in general suggests that the manner in which an object is segmented into parts depends on the learned informativeness of the parts.

On the basis of our results, we can ask whether categorization training improves the response to previously relevant parts, or impedes the processing of irrelevant parts. Evidence in favor of both processes was found in the experiments. In favor of training having a positive effect on relevant parts, it was found in Experiment 1 that relevant parts were identified as present or absent more quickly than either irrelevant parts, or complements of relevant parts.

There is also convincing evidence that training causes irrelevant parts to be ignored or rejected. In Experiment 1, participants were quicker to respond “absent” when a non-diagnostic feature was present in the whole object than when a complement was present. A similar bias to respond “absent” quickly was found in Experiment 2 when comparing nondiagnostic normal category parts to reflections of these same parts. Even more persuasive evidence that irrelevant features are processed less effectively comes from the comparison of normal parts and their reflections in Experiment 2. In Experiment 2, both “present” and “absent” judgments were slow for nondiagnostic parts relative to reflections of those parts. “Present” and “absent” judgments for diagnostic parts were roughly equal in speed to judgments about reflections of diagnostic parts. Thus, by comparing
judgments to their reflected controls, it becomes clear that one influence of categorization training is to desensitize irrelevant parts.

This desensitization of irrelevant parts is particularly surprising because it requires that the items should not be simply interpreted in terms of their diagnostic parts. Rather, the nondiagnostic parts must also be registered at some level in order for it to be inhibited. Although parsings of items into nondiagnostic and diagnostic parts are mutually inconsistent because they involve overlapping line segments, participants seem to generate both parsings. Rather than simply being ignored, nondiagnostic information seems to be actively suppressed. This conclusion is consistent with recent results showing that alternative figure-ground interpretations of a display compete against one another [Peterson and Lampignano (2003)].

Our current results complement other related studies showing the influences of category learning on the segmentation of objects. Hock, Webb and Cavedo (1987) showed that category learning increased the likelihood of segmenting a pattern into parts that were similar for patterns that were members of the same category. Finally, researchers have shown that participants’ ability to perform a figure-ground segmentation depends on their familiarity with the stimuli [Peterson and Gibson (1994), Vecera and O’Reilly (1998), Peterson and Lampignano (2003), Vecera et al. (2004)]. People’s lifelong familiarity with objects facilitates their ability to extract these objects from surrounding context and treat them as figures [Schyns and Murphy (1994)].

If our results are best explained by postulating that people create perceptual units for often-repeated patterns that are useful for categorization, one question that remains is, “How are these new units acquired?” Some researchers [Shiffrin and Lightfoot (1997), Goldstone (2000)] refer to a process of perceptual unitization by which conjunctions of stimulus features are “chunked” together so that they become perceived as a single whole unit. Simple co-occurrence of line segments is not sufficient for their unitization; nondiagnostic and diagnostic parts occur equally often during categorization. Within this framework, the sensitization of diagnostic over nondiagnostic features must be due to a unitization process that depends on categorical relevance as well as co-occurrence of features.

Mozer et al. (1992) have developed a connectionist model that learns how to segment objects. Their MAGIC system learns how to group features based on a set of presegmented examples. Object parts that belong to the same segment are represented in MAGIC by units that have the same phase of activation (they are firing in synchrony). Our experiments provide support for MAGIC’s flexible, rather than fixed, segmentation procedure. Mozer (1994) added a learning principle to MAGIC that does not require explicit feedback to be provided about part segmentations. In this new model, objects tend to be segmented into parts that are uniform across instances. According to his regularity principle, features within a natural part tend to have higher correlations in their structures than do features from different parts [for a similar principle, see Schyns and Murphy (1994)]. This newer approach shows even more promise of being able to account for our results because our categorization training does not provide explicit feedback about what segments should be bound together, but it does provide information
about co-occurrence relations between line segments. Again, in order to account for our experiments, this model would have to incorporate information about the categorization of objects, and not just relations between features within an object.

Goldstone (2003) presents a model of unitization, and the complementary process of differentiation, that does take into account the categorization of objects as well as unsupervised statistics across the entire set of objects. It possesses units that intervene between inputs and category outputs and can be interpreted as learned feature detectors. The CPLUS model is given a set of pictures as inputs, and produces a categorization of each picture as output. Along the way to this categorization, the model comes up with a description of how the picture is segmented into pieces. The segmentation that CPLUS creates will tend to involve parts that (1) obey the Gestalt laws of perceptual organization by connecting object parts that have similar locations and orientations, (2) occur frequently in the set of presented pictures, and (3) are diagnostic for the categorization. The network builds detectors at the same time as it builds connections between the detectors and categories. The psychological implication is that our perceptual systems do not have to be set in place before we start to use them. The concepts we need can and should influence the perceptual units we create.

3. Semantic reorganization during category learning

Several models of object perception have assumed that we recognize objects by compounding primitive elements such as features [Treisman and Gelade (1980)] or shapes [Biederman (1987)]. Likewise, many theories of conceptual representation have also been based on a fixed set of primitive semantic concepts [Schank (1972), Wierzbicka (1992)]. Just as we have favored approaches with adaptive perceptual elements, we have been led by our research to conclude that conceptual elements are similarly adaptive.

3.1. Integral versus separable dimensions

Our second line of research explores the flexibility of conceptual dimensions as they apply to classification. There has been a long history of research into how pairs of dimensions are processed, starting with Garner (1974, 1976) and Monahan and Lockhead (1977). Garner made the distinction between separable dimensions, for which one dimension can be attended to while the other is ignored, and integral dimensions, for which such selective attention is impossible. This distinction was based on patterns of results in classification tasks developed by Garner (1974). In the “correlated” task, values on both dimensions were varied together to form the stimulus set. For example, if the dimensions were size and shape of figures, then the correlated task would consist of large squares in one category and small circles in the other category. In the orthogonal (“filter”) task, the categorization rule depends on only one dimension and the other, irrelevant, dimension must be ignored. For example, figures might be categorized based on size (large vs. small) regardless of their shape (square vs. circle). Performance on these
tasks was compared to a univariate (“control”) task in which the stimuli were categorized on a single dimension with no variation in the irrelevant dimension.

In these tasks, one of two patterns often emerged for a given pair of dimensions. For integral dimensions (e.g., saturation and brightness), the correlated task was performed better than the control task, and the filter task was performed worse than the control task. For separable dimensions (e.g., size and brightness), the correlated and filter task performances were approximately equal to the performance on the control task. The degree of integrality of the stimuli was judged according to the amount of facilitation in the correlated task and the amount of interference by the irrelevant dimension in the filter task, as compared to the control task. The interference of the irrelevant dimension can be understood as the result of an inability to selectively attend to the relevant dimension. Likewise, the benefit of the redundant information in the correlated task could be due to both dimensions being used to perform the task, even though only one dimension is logically necessary. Monahan and Lockhead (1977) proposed that stimuli consisting of integral dimensions are initially processed in terms of overall similarity and then in terms of individual aspects. The reverse may be true for separable dimensions.

King, Gruenwald and Lockhead (1978) studied performance on the Garner classification tasks for animal terms based on the dimensions of size and ferocity. They found that the correlated task was performed better than the control task, which was performed better than the filter task. They interpreted the pattern of results as an indication of integral dimensions.

3.2. Experiment 3

To investigate the effects of category training on the integrality of semantic dimensions such as those used by King et al. (1978), we used a training-transfer paradigm using the correlated, filter, and conjunctive classification tasks. As Figure 7 shows, in the correlated task, either dimension can be used to perform the classification, or both can. In the filter task, only one dimension is relevant and the other dimension is irrelevant. In the conjunctive task, both dimensions are necessary. We hypothesized that the correlated task would induce more integral processing of the semantic dimensions since a conjunction of values indicates the category membership and this should facilitate the use of both dimensions as a unified single dimension. The conjunctive task should also induce a more fused representation of the two dimensions since both dimension values must be attended to in order to make a category choice. In the filter task, only one dimension is relevant to the categorization, so we hypothesized that this should induce a more separate use of the two dimensions. We measured these effects by training participants on one task (correlated, filter, or conjunctive) and then transferring them to a different task (correlated, filter, or conjunctive). If category training can affect the integrality of semantic dimensions, then positive transfer should occur if participants are trained on an integrating task (correlated or conjunctive) and then transferred to the other integrating task. Negative transfer should occur if participants are trained on an integrating task (correlated or conjunctive) and then transferred to the separating task (filter) or vice versa.
3.2.1. Method

Three word sets of 40 words each were designed from the categories of animals, vehicles, and clothing. For each word set, two dimensions were used. Two values were designated for each dimension, and the two dimensions were crossed, resulting in four cells with ten words in each cell (see Table 1 for the vehicles example). The dimensions were ferocity and sociability, capacity and speed, and warmth and casualness for the animal, vehicle, and clothing word sets, respectively.

One hundred and sixteen participants performed a training task followed by a testing task for each of the three word sets. Each of the three tasks in training was paired with one of the two different tasks in testing, resulting in six training-testing conditions (see Figure 7). In the correlated task, the categorization rule was based on the combination of two values on the dimensions that varied together. Words from two diagonally positioned cells were shown, but not from the other two cells along the reverse diagonal. In the filter task, the categorization rule was based on a single dimension that divided the set into two categories with two cells in each category. In the conjunctive task, the categorization rule was based on the combination of two dimensions for
Category X (one cell), while the remaining three cells formed Category Y. Table 2 shows the categorization rules for the set of vehicle words. Before each task, participants were told the general category (e.g., vehicles), the rule for both categories (e.g., Table 2), and the list of words for each category listed in columns (e.g., Table 1).

The participants were given 160 trials in each of the three training tasks and 120 trials in each testing task. The training tasks were divided into four blocks of 40 trials each. The testing tasks were divided into three blocks of 40 trials each. For each block, the words were selected with equal frequency from each cell and presented in a randomized order.

On each trial, the word was presented on the computer screen with the first letter at the center of the screen. Participants made their category choice using the number keys. They were given feedback on their choice using a check mark for correct answers and an X for incorrect answers.

### 3.2.2. Results and discussion

The average response time results for correct trials are shown in Figure 8. During training, the correlated and conjunctive tasks were both performed significantly faster than the filter task. The correlated testing task was not performed significantly differently
based on the training task that preceded it. Performance on the conjunctive testing task was significantly more accurate when it was preceded by the correlated training as compared to the filter training.

The performance during training provides a baseline to which we can compare the relative effects of training on that task. The correlated training had no significant effects on the conjunctive testing task compared to initial conjunctive training performance. The filter training had a significant negative effect on the conjunctive testing task compared to the initial conjunctive training performance for accuracy and response time. The filter training also had a negative effect on the correlated task compared to the correlated training task performance that was significant in terms of response time. The correlated training had a significant negative effect on accuracy of the filter testing task compared to filter training accuracy. The conjunctive training also had a significant negative effect on the accuracy of the filter testing task compared to filter training accuracy.

The filter task training resulted in negative transfer to the conjunctive task and the correlated task. The correlated task training did not have any effect on transfer to the conjunctive task. Relative to initial performance, we have evidence of negative transfer of training on the filter task for both the conjunctive and correlated tasks and negative transfer of training on the correlated and conjunctive tasks for the filter task. This matches the prediction that there would be negative transfer effects between tasks inducing separation of dimensions and those inducing fusion of dimensions.

Fig. 8. Response time data from Experiment 3 for the initial training tasks and the testing tasks (error bars ±1 S.E.).
3.3. Experiment 4

Experiment 3 showed that the effects of classification task training on subsequent testing tasks are consistent with an adaptation of the conceptual dimensions. However, two possible types of adaptation could be taking place: a change in the representation of dimensions for individual word exemplars of the category, or a change in the representation of the dimension at the category level. In Experiment 4, this question is explored using a design in which new words are introduced during the test tasks.

The design of the tasks was the same as in Experiment 3, except that new exemplars were presented in the testing phase that had not been presented in the training phase. We hypothesized that training in a task that induces processing of two semantic dimensions in an integral manner (correlated and conjunctive) will result in positive transfer to the other fusion-inducing task. Conversely, negative transfer is expected from training in the task that is thought to induce separate processing of dimensions (filter) to the fusion-inducing tasks (correlated or conjunctive) and vice versa. We also hypothesized that the same pattern of results in the testing tasks would be found for the both the novel words and the words previously seen during training although the negative transfer effects may be more pronounced for words that had previously been seen.

3.3.1. Method

The materials were similar to those used in Experiment 3, except that the number of words in each domain was doubled to 80. The three domains were animals, activities, and things. The animal words were again placed in a 2 × 2 table along the dimensions of ferocity and sociability. The activities word set consisted of sports and hobbies that were classified according to how physical the activity is (‘strenuous’ vs. ‘light’) and the riskiness of the activity (‘risky’ vs. ‘safe’). The things word set consisted of various objects and materials that were classified according to their naturalness (‘natural’ vs. ‘artificial’) and their fluidity (‘solid’ vs. ‘fluid’). All of the categories were pretested by having participants rate words on two dimensions and using these ratings to select words that fell most clearly into one category or the other.

Each of the 248 participants was given only one of the combinations of training and testing tasks. They repeated the particular train-test condition for each of the three word sets. All other aspects of the task design were the same as in Experiment 3 except for the use of new words in the testing phase. In the training task, only half of the available words in each cell were presented (10 words). In the testing task, all the available words for the cells used in the task were presented. The category frequency was balanced in each task and the order of the word sets and word presentations was randomly selected for each participant.

3.3.2. Results and discussion

The response times for the experiment are shown in Figure 9. The correlated testing task was performed significantly slower when preceded by the filter training than by the
conjunctive training, but this effect was limited to previously trained words only. In the filter testing task, the previously trained words were performed more accurately when preceded by the correlated training than the conjunctive training. In the conjunctive testing task, the novel words were judged faster when preceded by correlated training than filter training. Likewise, old words were categorized more accurately and faster when preceded by correlated training than filter training.

As in Experiment 3, the training task performances were compared using the average performance over each of the three blocks. And again, the correlated task was performed more accurately and faster than the conjunctive task. In turn, the conjunctive task was performed more accurately and faster than the filter task. Thus, for the initial training task performance, the same pattern of results was found as in Experiment 3: the correlated task elicited the best performance followed by the conjunctive task, followed by the filter task.

Novelty of the words during the testing task was the crucial factor tested in Experiment 4. Overall, words previously seen in training were responded to more quickly than new words and there was a larger range of improvement for the speed of response to new words than old words. These effects are not surprising since the old words presumably were made easier to classify in the testing task by the prior exposure in the training task. Therefore, more improvement is expected in the new words.

In the interactions between novelty and condition, there was little difference between the old and new words. For novel words, the effect of transfer is based on a shift in the integrality or separability of the semantic dimensions alone, and not due to a direct change in the specific item representation since the words were not present in training. The degree to which the training induced a change in the dimension representation over
and above the changes to individual item representations can be measured by the degree to which the same pattern of results is found for both old and new words. The correlated testing task revealed a benefit in terms of accuracy from the conjunctive training task compared to filter training for old words, but not for new words. The conjunctive task exhibited a positive transfer effect on old words from the correlated training compared to filter training, in terms of both accuracy and response time, and on new words in terms of response time. These results suggest that changes may occur in both the item representations and the semantic dimensions.

4. Conclusions on semantic reorganization

Both of the experiments in this series obtained the same result for the initial task performances. The correlated task was performed the best, followed by the conjunctive task, and the filter task was performed least well. The fact that both the correlated and conjunctive tasks were performed better than the filter task is likely the result of the dimensions’ initially being integrally processed such that they can be easily processed together but not so easily processed separately. These results echo the findings of King et al. (1978).

Negative transfer effects were obtained in the filter task following conjunctive or correlated training. Likewise, negative transfer effects were obtained in the correlated and conjunctive tasks following filter training. These effects support the hypothesis that training may induce a change in the integrality of the semantic dimensions.

Experiment 4 tested whether the adaptation occurs on an individual linguistic concept level, whereby the features of a particular item become integrated, or on a semantic dimension level, whereby changes generalize to other concepts defined by the altered dimensions. While some effects did not generalize to concepts not seen during training, correlated training had a positive effect on the conjunctive testing task relative to the filter training effects for both old and new words, suggesting that changes took place at the level of the semantic dimensions.

Experience using semantic dimensions in classification tasks can alter the processing of those dimensions. There were shifts in the apparent integrality of the dimensions such that tasks that incorporate both dimensions together may create a more fused representation of the dimensions. Other tasks that require the use of a single dimension and the discounting of an irrelevant dimension tend to cause a separated representation of the dimensions. More generally, our studies provide a methodological tool for examining how any number of semantic dimensions across domains are processed and adapted during classification tasks.

5. Integrating perceptual and semantic reorganization

Together, the four reported experiments suggest an alternative approach to theories that have posited fixed sets of perceptual [Treisman and Gelade (1980), Biederman (1987)] or semantic [Schank (1972), Wierzbicka (1992)] features. According to this alternative,
category learning not only uses existing object descriptions, but also alters object descriptions to facilitate the learning. Understandably, the claim that new perceptual or semantic features are created during category learning has been controversial [Schyns, Goldstone and Thibaut (1998)], and we would like to dispel the notion that feature creation is a magical process, or that once feature creation is allowed, then “anything goes.”

5.1. Characterizing psychological features

To understand what we mean by feature creation, it is helpful to first analyze what we mean by “feature.” By “feature” we mean a psychological unit of perception or thought. Dimensions are also psychological entities, but refer to a set of values that can be ordinally positioned. Brightness, then, is a psychological dimension only because it is processed as a unit. If luminance energy were not psychologically isolated, then there would not be a (psychological) dimension of brightness reflecting this physical quantity.

If features and dimensions are units of perception and thought, then we can ask what physical aspects are bundled together into these psychological units. Features can be interpreted as packages of stimulus elements that are separated from other sets of elements and that reflect the subjective organization of the whole stimulus as a set of components. Features can be revealed using several experimental operationalizations. If two pieces of physical information, X and Y, are packaged together in the same psychological feature and Z is not, then several empirical predictions follow. We predict that searching for X and Y simultaneously should be easier than simultaneously searching for X and Z [Treisman and Gelade (1980)]. We predict that searching for X should be affected by contextual variation to Y more than to Z [Gauthier and Tarr (2002)]. And we predict that categorization based on X should be slowed more by irrelevant variation to Y than to Z [Garner (1974, 1976)]. It should be easier for people to attend to X and Y simultaneously than X and Z. All of these operationalizations tie into the notion that X and Y are being processed together.

It is also noteworthy that all of these operationalizations imply a continuum of featurehood. There will be various degrees to which stimulus aspect Y intrudes upon or facilitates processing of X. Although we conceive of features as packages of stimulus components, we are not proposing that packages are completely discrete or mutually exclusive. Rather, they are packages in the same way that cliques can be circled in social networks or regions can be identified in brain neural networks. In all three domains, a unit (feature, clique, or region) is characterized by relatively dense within-unit connectivity among elements and relatively sparse connectivity between elements within the unit and external elements. Features are useful idealizations because they capture the notion of elements that are densely interconnected, but it is important to recognize that (1) features (e.g., densely interconnected clusters) may exist at multiple levels of resolution, (2) elements processed as one feature may not have uniform interconnectivity to other elements of the same feature, and (3) the internal integrity of different features may vary.
5.2. Characterizing featural change

Having characterized psychological features, we can now turn to the meaning of feature creation. By this account, feature creation simply involves alterations to the organization of stimulus elements into features. Figure 10 shows two ways that this can happen. By unitization, stimulus elements (circles) that were originally processed as three different features (ovals) come, with practice, to be processed as only two features. Elements that were originally processed separately are now processed together [Shiffrin and Lightfoot (1997), Goldstone (2000)]. By differentiation, the same three-element object comes to be processed as four features. Elements that were originally psychologically fused together become isolated [Smith and Kemler (1978), Smith, Gasser and Sandhofer (1997), Goldstone and Steyvers (2001)].

From Figure 10 it may appear like there are two separate, perhaps contradictory, tracks for featural change. In fact, not only are unitization and differentiation compatible with each other, but they often occur simultaneously. They are compatible because both processes created appropriate-sized units for a task. If elements covary together and their co-occurrence predicts an important categorization, then the elements will tend to be unitized. If elements vary independently of one another and they are differentially relevant for categorizations, then the elements will tend to be differentiated. Experiments 1 and 2 are good examples of simultaneous unitization and differentiation. During category learning, the three line segments that jointly indicate a category are unitized together, and are isolated from other line segments in the objects to be categorized. Accordingly, we do not support theories that propose monolithic developmental trends toward either increasingly unitized [Gauthier and Tarr (2002)] or differentiated [Kemler and Smith 1978] representations. We believe that both occur, and furthermore, that the same learning algorithm can do both simultaneously [Goldstone (2003)].

Fig. 10. Two varieties of featural reorganization. Stimulus elements are shown by circles and psychological packages of those elements – in features – are shown by oval. By unitization, stimulus elements that were once processed as different features come to be processed as a single feature. By differentiation, stimulus elements that were once processed as the same feature come to be processed by multiple features.
Features are not created “out of nothing.” They are reorganizations of stimulus elements. A critic might respond, “Then how is your account any different from the standard fixed-features approach in which primitive elements are combined in new arrangements to create object representations?” For now, we will give three replies [see Schyns et al. (1998) for others]. First, by our account, features are not (always) created from a set of psychological primitives. Often they are created from stimulus elements that originally have no parsing in terms of psychological primitives. For example, people can create a “saturation” detector that is relatively uninfluenced by brightness even if there was originally no detector that had this response profile [Burns and Shepp (1988)]. To be sure, if brightness and saturation affected a brain identically, then there would be no way to develop a detector that responded to only one of these properties. However, as long as two properties have some differential effects, then increasingly differentiated detectors can emerge, if the training encourages their isolation. The critic might counter, “But dimensions that are fused together at some point in perceptual processing can never be split later.” By analogy, once red ink has been poured into blue ink, there is no simple procedure for later isolating the blue ink. Fortunately, this analogy is misleading, and there are several computational models that can differentiate fused dimensions [Smith et al. (1997), Edelman (1999), Goldstone (2003)]. For example, competitive learning networks differentiate inputs into categories by developing specialized detectors for classes of stimuli [Rumelhart and Zipser (1985)]. Random detectors that are slightly more similar to an input than other detectors will learn to adapt themselves toward the input and will inhibit other detectors from doing so. The end result is that originally homogeneous detectors become differentiated and heterogeneous over the course of training.

Second, feature creation often involves delineating spatial regions rather than composing elements. For example, a bounded segment of a curve can be extracted by identifying its end points by rapid changes in curvature [Hoffman and Richards (1984)]. This extraction does not require piecing together elements. What would these putative elements be – line segments or pixels? There is good evidence that neither small line segments nor pixels are functionally useful features for object recognition. They are too low-level to provide diagnostic evidence for actual objects. Moreover, pixels cannot be true features because they are not identified by intrinsic attributes like ‘red’ or ‘4 cm.’ Their essential nature depends upon their location in spatial media. Much of feature creation involves forming bounded regions in a spatial medium rather than symbolically composing atomic elements.

Third, there are clear-cut cases, where something like new perceptual devices are created. By becoming physically modified, systems can learn to represent properties that they were unable to represent originally. In evolutionary time, organisms developed ears sensitive to acoustic properties that no early organisms (e.g., bacteria) could detect. This is also possible within a system’s own lifetime. The cybernetician Gordon Pask built a device that could create its own primitive feature detectors. It consisted of an array of electrodes partially immersed in an aqueous solution of metallic salts. Passing a current through the electrodes caused dendritic metallic threads to grow. Eventually
the threads created bridges between the electrodes, which subsequently changed the behavioral repertoire of the device. Cariani (1993) reports that within a half a day, the system could grow to be sensitive to a sound or magnetic field. With more time, the device could discriminate between two musical pitches. Similarly, there is good neuro-physiological evidence that training can produce changes in the early somatosensory, visual, and auditory cortex [see Goldstone (1998) for a review]. While these changes are not as radical as sprouting a new ear, they are existence proofs that early perceptual devices can be systematically and physically altered by the environment to change their representational capacities.

5.3. Prospects for synthesizing perceptual and semantic reorganization

We have juxtaposed two series of experiments with the intention of highlighting similarities and differences between the perceptual and semantic reorganization that accompanies concept learning. In the first series of experiments, people apparently create shape complexes during category learning, and use those shape complexes as building blocks for describing subsequently presented objects. In the second series, people create either fused or separated semantic descriptions that subsequently affect their later categorizations. Is the process of creating a three-line-segment complex similar to creating an integrated representation of the timidity and sociability of animals, or the speed and capacity of vehicles?

One apparent discrepancy between perceptual and semantic unit construction is that there are strong visuospatial constraints on perceptual unit creation. People have a strong bias to create units that obey Gestalt laws of proximity, similarity and good continuation. These biases are needed for computational models that aim to create psychologically plausible perceptual units [Goldstone (2003)], and are useful in limiting the combinatorial explosion of potential units that could be built. At first sight, semantic units do not have any corresponding constraints on their construction.

Upon further reflection, we believe that there are biases affecting semantic unit construction, and that these biases play a loosely analogous role to the Gestalt laws of perceptual organization. Informal interviews with some of the participants in Experiments 3 and 4 suggest that in the conjunctive and correlated conditions, participants often created conceptions that fused the two component dimensions into a semantic Gestalt. For example, for the animals category, people often created a schema for social, timid animals that consisted of groups of small animals huddled together for protection. For the vehicles category, participants sometimes created a fused notion of fast, high-capacity vehicles by imagining mass transportation systems. By this account, just as it would be difficult to create a unit for two line segments that are far apart, of different thicknesses, and not part of a continuous path, it should be difficult to create semantic units for hard-to-relate semantic dimensions such as Jorge Luis Borges’ (1966) dimensions of animals: “those that tremble as if they were mad” and “those that have just broken a flower vase.” Furthermore, we believe that these constraints on semantic unit construction are important for creating nontrivial units. There is a trivial sense in which any features,
such ‘square’ and ‘blue,’ can be combined to create a conjunctive unit ‘square and blue.’ However, these conjunctions are inert, being no more than the Boolean concatenation of their elements. For these conjunctions, the standard compositional account of unit construction is perfectly adequate. However, semantic reorganization often differs from logical combination, and the elements interact to create complexes with emergent properties. Much of the recent work on knowledge-based categorization provides insight into the development of semantic complexes [Murphy (2002)].

Much of the most important work in characterizing representational reorganization will involve specifying mechanisms that are tightly tied to particular classes of materials. Still, we are sanguine about the heuristic utility of attempting to unify perceptual and semantic reorganization processes. Complementary mechanisms of differentiation and unitization are found for both. Both are guided by unsupervised statistics and supervised feedback provided by categorizations. Moreover, it may prove difficult to draw a clean dividing line between perceptual and conceptual processing [Goldstone and Barsalou (1998)], not just because we lack precise enough empirical diagnostics, but because they emanate from a shared substratum.

Acknowledgments

Many useful comments and suggestions were provided by John Kruschke, Robert Nosofsky, Linda Smith, and Richard Shiffrin. This research was funded by NIH Grant MH56871, and NSF Grant 0125287.

References


Chapter 30

THE RETURN OF CONCEPT EMPIRICISM

JESSE J. PRINZ

Department of Philosophy, University of North Carolina at Chapel Hill

Contents

Abstract 679
1. Introduction 680
2. Concept empiricism 680
   2.1. Representing and doing: Two faces of concepts 680
   2.2. Variable mechanisms 684
   2.3. Perceptual vehicles 686
   2.4. Innateness 688
   2.5. Summary 691
3. The abstract ideas objections 692
References 694

Abstract

In this chapter, I outline and defend a version of concept empiricism. The theory has four central tenets: concepts represent categories by reliable causal relations to category instances; conceptual representations of categories vary from occasion to occasion; these representations are perceptually based; and these representations are all learned, not innate. The last two tenets on this list have been central to empiricism historically, while the first two have been developed in more recent years. I look at each in turn, and then I discuss the most obvious objection to empiricism: that some concepts cannot be perceptually based because they represent things that are abstract, and hence unperceivable. I discuss two standard examples: democracy and moral badness. I argue that both can be explained using resources available to the empiricist.
1. Introduction

The history of Western philosophy can be viewed as a debate between rationalists and empiricists. Rationalists emphasize innate concepts, the power of *a priori* reasoning, and the unreliability of perception. Empiricists regard perception as the source of our concepts and the primary means of attaining knowledge. Since Plato and Aristotle, the pendulum has been swinging back and forth between these positions. The high point in this debate occurred in the seventeenth and eighteenth centuries, when continental rationalists, such as Descartes and Leibniz, revamped rationalism, and British empiricists, such as Locke and Hume, worked out the empiricist alternative. Cognitive science was born as the pendulum swung back toward rationalism. Before the cognitive revolution, Skinner had allied himself with the empiricists, and Chomsky launched a self-consciously rationalist assault on Skinner’s research program. Within psychology, the rationalist trend has continued. For example, a sizable proportion of research on concept acquisition focuses on innate domains of knowledge. There are, however, also signs of dissent. Connectionism and situated cognition have attracted considerable interest over the last decade, and both approaches depart from rationalism in a significant way. There has also been a revival of more traditionally empiricist theories. Of special interest is the work of Larry Barsalou and his colleagues. His group has, more than any other, carried the torch for the likes of Locke and Hume. Barsalou’s work has been my point of departure.

In this chapter, I argue that a traditional brand of empiricism has a lot to offer. I think it is time for the pendulum to swing back again. I will begin my case by presenting the core components of the empiricist theory that I favor. The most central tenet is that concepts have their basis in perception. A second tenet, which was equally important to Locke and Hume, is that concepts are learned. To these classical tenets I add two more contemporary suggests: concepts refer to categories in the world via reliably causal relations and concepts are contextually variable. I will discuss these four tenets in reverse order. My discussion of the first two is primarily directed at philosophers. My discussion of the second two will be of equal concern to psychologists, who may wish to skip ahead to Section 2.3. Together, I think these four tenets add up to a defensible version of empiricism. There is, however, one serious objection, stemming from the tenet that concepts are perceptually based. This is plausible in the case of concrete categories, but we are also capable of thinking about abstract things. It is not clear how perceptually based concepts can handle abstract thought. I conclude with a discussion of this objection.

2. Concept empiricism

2.1. Representing and doing: Two faces of concepts

In a recent paper, Jerry Fodor (2004) has suggested that proponents of a rationalist approach to concepts can be distinguished from defenders of other theories by their
view of what concepts are for. Rationalists claim that concepts are for thinking. More precisely, to have a concept is to be able to think about something. Having the concept [DOG] means being able to think about dogs. A better word for thinking, in this context, might be representing. Concepts are primarily in the business of representing. For opponents of rationalism, including empiricists, having a concept is being able to do something. For example, having the concept [DOG] might be construed as the ability to categorize or interact with dogs. Fodor’s way of setting things up distinguishes two broad functions for concepts: rationalists say that concepts are primarily in the business of representing, and opponents of rationalism say that concepts are primarily in the business of doing. This distinction should not be regarded as a disjoint dichotomy. Empiricists do not deny that concepts represent. Rather, they claim that concepts have other equally important functions. Empiricists say that concepts must be able to represent things in a way that facilitates interaction with those things. Representing must be in the service of doing.

The difference between Fodor and the empiricists is captured by a difference in the nature of the mental representations that they postulate. Fodor thinks concepts are like words. They are arbitrary symbols in a language of thought. There is little one can do with an arbitrary symbol. A symbol does not include any instructions for how to interact with the category that it represents. For the empiricist, concepts are more like mental images, or inner models. Representations of that kind can be used to guide action. We can read features of a category from our concepts if empiricism is correct.

I will discuss these issues of representational format below. In this section, I want to consider another question about the distinction between representing and doing. Fodor attempts to separate these two functions by developing a theory of representation that would allow concepts to represent things without encoding the kinds of features that would allow them to do anything. That theory constitutes one of the best current explanations of how concepts represent. If the theory supports Fodor’s rationalism, then empiricists have reason for concern. My goal here is to show that Fodor’s theory of reference is actually better suited to empiricism.

Fodor (1990) develops a theory of representation that promises to explain how concepts represent without making any mention of what they do. This allows him to secure the conclusion that representing is prior to doing, which is a central tenet of rationalism. I begin with a summary of this theory. According to Fodor, a concept represents that which would reliably cause the concept to be activated. A concept represents dogs if encounters with dogs would ordinarily cause that concept to activate. This is only a first approximation. Formulated in this way, the account faces an obvious objection. Our [DOG] concepts are activated when we encounter dogs, but they are also activated when we encounter things that merely look like dogs, e.g., foxes in bad lighting. If concepts represent anything that activates them, any concept that represents dogs would also represent foxes. To solve this problem, Fodor notes that there is an asymmetric dependency relation between dogs and foxes with respect to our [DOG] concepts: foxes would not cause our [DOG] concepts to activate were it not for the fact that dogs do, but the converse is not true: the fact that dogs cause our [DOG] concepts to activate is
not a consequence of the fact that foxes do. Fodor construes this asymmetry synchronically, in terms of counterfactual dependencies. I will not argue the point here, but I think the best way to make sense of it is diachronically. A [DOG] concept is one that was created in the context of dog encounters. Fox encounters would not cause the concept to activate were it not for dogs having done so in the past, but not conversely. On this reading, a concept represents a category when two conditions are met:

Nomological causation: the concept is disposed to be reliably activated by encounters with members of the category, and

Etiological causation: encounters with members of the category played a role in the acquisition of the concept.

Fodor thinks that concepts represent in roughly this way (with a synchronic condition in place of the etiological causation clause). He also thinks that this story favors the hypothesis that concepts are primarily in the business of representing, not doing. To see why, it is important to consider two other alternatives to this causal theory of reference. According to one view, concepts refer by resemblance. They are mental images that are structurally isomorphic with the things they represent. According to another view, concepts are feature sets that refer via description. The concept [DOG] refers to dogs, because the concept [DOG] contains a collection of features describing dogs, and dogs are the only things that satisfy that description. [DOG] contains [FURRY], [BARKS], [QUADRUPEDAL], and so on. By denying these two theories of reference, Fodor is able to defend the view that concepts are unstructured arbitrary symbols [Fodor (1998)]. They are words in a language of thought. Fodor can defend the language of thought story only by arguing that concepts do not depend on description or resemblance to refer. An individual word does not describe anything, and it does not look like what it refers to. By embracing a causal theory of reference, Fodor explains how word-like mental representations can refer. Without this, it would be difficult to maintain that concepts are couched in an arbitrary code. It would also be hard to maintain that concepts are primarily in the business of representing. An arbitrary symbol cannot be used on its own to recognize dogs or draw inferences about dogs. It is a dog symbol in the purest sense: it represents dogs and does nothing else. A mental image of a dog represents dogs and can also be used to recognize them. A dog description represents dogs and can also be used to draw inferences about them. Fodor's causal theory of reference secures his hypothesis that concepts are primarily in the business of representing, not doing.

Fodor's mental word theory is radically different from the way most psychologists think about concepts. Psychologists emphasize the role that concepts play in categorization. If concepts are tools for categorizing, they cannot be unstructured word-like entities. They must be built up from features. Some psychologists say that [DOG] is a prototype, others say it is a mini-theory, and still others say it is a set of exemplar representations. As a rationalist, Fodor thinks that a theory of concepts need not explain how we categorize. His mental word theory is ideally suited for rationalism. After all, words in public languages represent, but we cannot categorize with them; they are arbitrary symbols. In Fodor's view, categorization is achieved by independent mechanisms.
He does not offer an account, but he might say that we have complex mental databases containing perceptual information, theories, prototypes, memory traces, and any number of features and facts. That is to say, categorization is achieved using the kinds of mental mechanisms that psychologists postulate. Think of a concept as a label on a large mental file. Information in that file and information stored elsewhere can play a role in categorization. The concept hovers safely above the overflowing sheets and scraps in the file. Items in the file represent what category members look like, the ontological domain they belong to, the attributes of specific instances, and so forth. But only the label represents the category itself.

On the face of it, Fodor seems to have what he wants. He has a theory of how mental representation works that is consistent with rationalism. Concepts are arbitrary symbols that can be used for nothing other than representing categories. They cannot be used to draw inferences, to plan actions, or to categorize. All of those functions are handled by the contents of our mental files. But this picture is very odd. It renders concepts needlessly anemic. Why should we say that concepts are arbitrary *labels*, rather than identifying concepts with the *contents* of our mental files? After all, the contents of those files do much more work. They allow us to categorize and act. Moreover, these files are absolutely essential for Fodor’s own theory of representation. A mental label represents a category by being reliably activated by instances of that category. But the label can be activated by category instances only if we have mechanisms that allow us to recognize those instances. An arbitrary [DOG] symbol can be triggered by dogs only if we have resources for recognizing dogs. Fodor assumes that all of the necessary resources are contained in our mental files. But once he makes that concession, the arbitrary labels begin to look unnecessary. It seems we should identify concepts with the file contents, rather than the file labels. We should say that concepts are the mechanisms that allow us to recognize categories rather than arbitrary mental words that flash on in the head when a category has been recognized. The labels are entirely unnecessary.

The moral is that Fodor’s theory of representation may not favor rationalism after all. Once we adopt a causal theory, we are forced to postulate mechanisms that allow us to reliably detect category instances. Once we postulate such mechanisms, we might as well identify them with concepts. We do not need to postulate arbitrary labels. Concepts can be complex databases. Such databases allow us to represent, but they also allow us to do things: they allow us to interact successfully with the world. So representing and doing are not disjoint functions, on this picture. They are intimately linked. According to Fodor’s theory, we think using unstructured symbols. [DOG] is just an arbitrary word in the language of thought. In the view I am recommending, [DOG] is constituted instead by the representations used to identify dogs. Thus, [DOG] is constituted by features that tell us what dogs look like and what their behavioral dispositions and affordances for interaction are. These features allow us to represent dogs, by securing reliable causal relations with them, but they also allow us to recognize dogs, and do things with dogs.

Rather than saying that concepts are for representing, I would say that representations are for doing. The only reason we represent the world is to make our way through
it. If concepts are not guides to possibilities for action, they are not useful. Concepts that merely represent belong to the fictional realm of pure Cartesian egos. Conceptual capacities that evolved in the real world allow us to run for cover or play fetch. But concepts also represent. The mechanisms that allow us to identify objects and interact with them also, thereby, establish reliably causal relations with those objects. Fodor himself shows how such causal relations can be used to establish reference. Ironically, his theory of reference fits perfectly with the antirationalist program.

2.2. Variable mechanisms

Fodor has reasons for identifying concepts with arbitrary labels, rather than the contents of mental files. He did not come to this conclusion without any arguments. He thinks that all theories that identify concepts with complex data structures, rather than unstructured word-like entities, are hopeless. If he is right, concepts cannot be identified with the complex mechanisms in our mental files that we use to identify members of a category. He has three main reasons for this conclusion:

1. The mechanisms in question do not combine compositionally. Concepts must be compositional, because that is the only way to explain our capacity to continually generate new thoughts. Therefore, concepts cannot be the mechanisms of categorization.

2. The mechanisms in question are highly varied and highly variable. Thus, they are unlikely to be the same from person to person or time to time. Concepts are also used to assign meanings to words. People must assign the same meanings. Otherwise, they would not be able to communicate. So concepts cannot be very variable, and they cannot be the mechanisms of categorization.

3. Concepts represent via reliable causal correlations with the things they represent. If concepts were as variable as the mechanisms that we use to identify category instances, then they would not be highly correlated with the categories they represent. Different concepts would be used for different encounters with the same category. That would prevent concepts from serving as representations of those categories. Therefore, concepts cannot be mechanisms of categorization.

I discuss the first two problems elsewhere [Prinz (2002), Prinz and Clark (2004)]. I want to turn my attention to the third—the claim that representation requires conceptual invariance. This claim is only implicit in Fodor, but I think it is the best argument for the view that concepts are like words.

Let us assume, with Fodor, that categorization cannot be achieved using invariant representations. We have mental files filled with different kinds of representations that are recruited on different occasions. We might even generate new representations for a category on the fly, by combining information in the file for that category with other information that pertains to the present context. The question is, can such varied representations represent the category, as opposed to representing transient features of the category (e.g., its appearance in this instance right now)? I think the answer is yes. To make this point, I need to show that the theory of representation sketched earlier is
applicable to variable representations. If variable representations satisfy the two conditions in that theory (etiological and nomological causation), then variability presents no barrier to explaining how concepts represent.

Let us begin with the etiological condition. According to that condition, a mental representation can represent a category only if the representation is an instance of a type of representation that was acquired as the result of an encounter with that category. This condition is satisfied by the representations that we use in categorization even if those representations are variable. The highly varied representations used to categorize instances of a particular category have something in common: they derive from the same mental file. That file was established when we first encountered instances of the category in question. Every time we encounter an object that matches information already contained in the file, we have an opportunity to add new information. We can also expand files by reflecting on the information that they already contain. Because this information is bundled together in memory, we are able to do the requisite bookkeeping. If we represent a category using item A from a file on one occasion, and item B from the same file on another occasion, we know that A and B are culled from the same source. In this sense, the heterogeneous representations used to represent a category on different occasions satisfy the etiological causation condition on representation, which I introduced earlier. All of the items in a mental file trace back to the time when the file was first created, and none of the items in that file would be there if it were not for the initial representations formed when we first encountered an instance of the category. Suppose my [DOG] file was created when I first saw a dog, and suppose that dog was a golden retriever. Later in life, I see a Pomeranian for the first time, and it looks similar to the retriever in certain respects, so it ends up in my [DOG] file. Now, one afternoon, I see a fox, which looks a lot like a Pomeranian, so it triggers the Pomeranian representation. Does this entail that my [DOG] file represents dogs and foxes? No. When I call a fox a dog, I am making a mistake. The [DOG] file was created as the result of an encounter with a dog. Other animals may happen to activate items in the file, but, since the file itself traces back to dogs, these are cases of misrepresentation. I conclude that variable representations can satisfy the etiological condition on representation.

Variable representations also satisfy the nomological causation condition, at least when they are considered collectively. There is a reliable causal relationship between encounters with members of a category and representations derived from the file for that category. My Pomeranian representation is not reliably caused by dogs, in general. It might not be activated when I encounter a sheepdog. But my Pomeranian representation is a member of a mental file containing variable dog representations, and collectively these are reliably caused by dogs. In other words, dog encounters reliably cause us to access the [DOG] file. Items in the [DOG] file are, in that sense, under the nomological control of dogs. Items in a mental file can be said to refer to the category that the file reliably detects. Thus, the nomological condition on reference can be met, even if the representations we use to categorize dogs are highly variable. I conclude that variability poses no barrier to representation. Concepts can represent categories even if they are instantiated in a variety of different ways.
In the view I am considering, the same concept will be constituted by different representations on different occasions. When you see a small long-haired dog, you will use one dog representation, and when you see a large short-haired dog, you may use another. In this sense, concepts vary with context. Barsalou (1987) reviews evidence that favors this conclusion. There is reason to think that we represent the same category in a variety of different ways. A typical dog in a French restaurant will differ from a typical dog in the arctic tundra. Barsalou shows that judgments of typicality vary as a function of imagined contexts. He concludes that concepts are not small sets of fixed features, much less unstructured words in a language of thought. Rather, concepts are temporary and variable constructions in working memory. They are drawn up in task-sensitive ways from large data structures in long-term memory.

2.3. Perceptual vehicles

The story, thus far, has two core tenets. One, borrowed from Fodor, is that concepts are representations of categories, and they represent, in part, by being reliably caused by category instances. The other, borrowed from Barsalou, is that concepts are highly variable constructions in working memory. The two tenets go together naturally. The variability ensures that concepts can be reliably activated by encounters with category instances. There is a third core tenet of the view that I favor, which is also found in Barsalou (1999): concepts are perceptually based.

To say that concepts are perceptually based is to say that they are made up from representations that are indigenous to the senses. Concepts are not couched in an amodal code. Their features are visual, auditory, olfactory, motoric, and so on. They are multi-media presentations. This tenet lies as the heart of classical British empiricism. Hume (1748, p. 62) says, “All our ideas are nothing but copies of our impressions.” I call this the modal specificity hypothesis.

The evidence for modal specificity comes from a variety of sources. Barsalou (1999) summarizes findings from psychology, neuroscience, and linguistics that are consistent with the idea that we think in perceptual codes. Damasio (1989) has argued, on the basis of functional neuroanatomy and neurological deficits, that thinking involves reactivating the perception centers that would be active if we were perceiving the things we were thinking about. Barsalou et al. (1999) have shown that people spontaneously use imagery in cognitive tasks. Lakoff (1987) has shown that we make pervasive use of perceptual metaphors when we verbally describe abstract domains.

Barsalou (this volume) offers an up-to-date review of findings that support the claim that concepts are perceptually based. I will not repeat his survey here, but let me briefly mention two of representative results. First, consider a study by Borghi, Glenberg and Kaschak (in press). If concepts are constituted by amodal symbols, then the features related to a concept should be organized as a list or a semantic network. Proximity in a semantic network should be based on semantic relatedness (e.g., dimensions and attributes should be directly linked) or strength of association (co-instantiated lexical items should be linked). If concepts are constituted by perceptual representations, then further
factors will contribute to feature organization, such as salience and special proximity. The perceptual theory of concepts predicts that features that are close to each other on a category instance will be close to each other in our mental representation of that category instance. Borghi et al. tested this by giving subjects a property verification task. Subjects first read sentences such as “You are washing a car.” Then they had to answer questions such as “Do cars have trunks?” or “Do cars have steering wheels?” Both features are strongly associated with cars, but subjects who were given the sentence about car washing were faster to answer the question about trunks. These results reversed when subjects began with the sentence “You are driving a car.” These results show two things. First, they reconfirm that representations are variable. The way we think about cars depends on the context (driving vs. washing). Second, our representations of categories are spatially organized; parts that are farther from our current perspective take more time to access, even if those parts are strongly associated with the category. This is predicted by the perceptually based theory of concepts and not by the amodal theory.

As a second example, consider a recent study by Pecher et al. (2003). They asked subjects to answer a series of property verification sentences. For example, subjects might be asked: “Are blenders loud?” “Are cranberries tart?” “Do leaves rustle?” The questions involved familiar features of objects. If concepts were represented in an amodal code, then features such as ‘loud,’ ‘rustle,’ and ‘tart’ would be represented in a similar way. If concepts are represented using modality-specific features, then the first two are stored in one sensory modality (audition) and the other is represented in another sensory modality (gustation). Thus, the perceptually based theory predicts that when subjects move from a question about loud blenders to a question about rustling leaves, they should be faster than if they had to move from a question about loud blenders to a question about tart cranberries. Shifting modalities should incur switching costs, if concepts are coded in a modality-specific way. This is just what Pecher et al. find. The task shows not only that perceptual features are associated with our concepts, but that we use such features when performing conceptual tasks. Defenders of amodal features do not predict this.

These findings provide empirical support for the hypothesis that concepts are perceptually based. Further support comes from theoretical considerations. Above, I endorsed the view that concepts represent by being reliably caused by category instances. This theory of representation is very popular in philosophical circles, and it has been defended by rationalists, such as Fodor. But the theory can actually be used to argue for the modal specificity of thought. If concepts are reliably caused by the categories that they represent, then every concept must be associated with a collection of perceptual features. Objects out there in the world can cause mental events only by impinging on our senses. So anyone who buys into the theory of representation presented earlier is committed to the view that we can perceptually identify category instances and that doing so is essential to representation. If perceptual states are essential for getting concepts to represent, we can simply hypothesize that concepts are copies of those perceptual states. This is just a minor addition to an argument already presented. I argued that concepts are the mechanisms by which we categorize. To that,
I now add that those mechanisms must be realizable in perceptual media, because categorization requires identification of perceived objects. Empiricism sounds radical at first, but it is actually consistent with the simple and obvious point that perception is needed to apply our concepts to things in the world. Every serious theory of concepts is committed to that. Empiricists take this universally accepted principle and run with it. If concepts are associated with perceptual representations, perhaps that is all we need. Postulating a further class of representations (amodal symbols) is unnecessary.

The modal specificity hypothesis is compatible with a parsimonious theory of how the brain evolved [see also Churchland (1986), Barsalou (1999)]. At first, creatures were input–output machines. The world caused sensory stimulation, and sensory stimulation caused programmed responses. Consider how flies avoid swatters, and you will get the idea. Over time, creatures evolved the capacity to store perceptual records of objects that they had encountered in the past, as well as the past consequences of those encounters. This allowed for much creative flexibility of response. Such creatures can respond differently to different objects even if they were not hardwired to recognize those objects. Finally, creatures evolved the capacity to reactivate the stored perceptual records in the absence of sensory stimulation, and the capacity to manipulate those records in working memory. That is what we do, and we do it better than any other creature on Earth. This story does not require the evolution of any special amodal codes.

2.4. Innateness

The theory of concepts that I favor has one more tenet, which also derives from the empiricist tradition. Locke (1690) began his case for empiricism by arguing against the doctrine of innate ideas. He believed that there were no innate concepts or principles. Empiricism was motivated by the idea that the attainment of concepts must involve learning, rather than triggering innate knowledge. These days, nativism about concepts is very popular. The majority of researchers working on concepts believe that we have quite a bit of innate machinery. The most popular suggestion is that we have innate knowledge of basic ontological domains, such as macro-object physics, biology, and psychology.

I do not believe that any of these domains is innate. That is to say, I do not think we have innate domain-specific knowledge that contributes to structuring our concepts. I cannot adequately defend the claim here. What I will offer instead is a few brief remarks and pointers. I hope that such an abbreviated discussion can at least serve to motivate hard reflection on the widespread nativist dogma. The innateness of core domains remains an open question for research.

First consider folk physics. Spelke (1994) argues that our understanding of certain physical principles must be innate because we seem to understand these principles early, whereas other principles, which are more perceptually salient, are not understood as early. For example, 3-month-olds seem to understand a principle of cohesion: objects move as connected wholes. When they habituate to what looks like a bar moving behind an occluding object, they dishabituate if the occluder is removed to reveal two bars.
moving in sync. They dishabituate less if the bar behind the occluder is solid, consistent with the adult expectation [Kellman and Spelke (1983)]. In contrast, infants at the same age do not understand gravity. They do not expect an object teetering far off the edge of a supporting surface to fall [Needham and Baillargeon (1993)]. Spelke thinks gravity is more obvious perceptually than object cohesion, so the latter must be innate.

Several things can be noted in response. First, expectations of cohesion can be learned by observation. When we see two shapes moving along the same spatiotemporal trajectory, they are almost always connected. Slater et al. (1990) showed that newborns do not make the coherent object assumption. Second, even if cohesion were not learned, there is little reason to think it is a conceptual capacity, rather than a feature of how our perceptual or attentional systems pick out objects [(Scholl and Leslie (1999)]. Third, gravity is often violated in an infant’s world, so it is not surprising that infants are slower to learn about it. Infants see hanging mobiles and doorknobs [Baillargeon, Kotovsky and Needham (1995)], and they have comparatively little experience with their own bodies falling. Finally, the Kellman and Spelke results may be an artifact of display complexity [compare Bogartz, Shinskey and Speaker (1997)]. It is widely known that infants will, at baseline, often stare at two objects for a longer time than they will stare at one object. In the Kellman and Spelke study, infants stared longer when an occluder was removed to reveal two bars rather than one. Perhaps they were staring longer simply because two objects are more interesting than one. Despite a variety of clever control conditions, Kellman and Spelke do not adequately rule out this hypothesis. For example, in one control condition, they habituated infants to a pair of moving bars with no occluder. At test time, infants stared longer at one bar rather than two. This shows that infants do not always stare longer at two bars. If the infants construed the display in the first experiment – the one with the occluder – as two bars rather than one, they should get bored of seeing two bars and show excitement when the occluder is removed to reveal a single bar. Since the infants in that experiment showed excitement about the two bars, Kellman and Spelke conclude that they construed the habituation display as a single bar. But this conclusion is too hasty. In the control condition, infants see two bars moving with a gap in between them. This looks just like the two bars in the test condition. In the version with the occluder, infants do not see two bars with an open gap between them. The occluder fills the gap. In other words, the appearance of the bars in the control condition is just like the appearance of the bars in the test condition, because there is an unfilled gap. In the occluder condition, the appearance changes, because a gap appears where there had been a surface. This difference may be enough to erase the effects of habituation and re-engage the infants attention. This suggestion is highly speculative, of course, but it is intended as a reminder that it is difficult to infer what infants are thinking from their performance in studies of this kind. When it looks as though infants understand a feature of the physical world, they may actually be showing their preference for perceptual similarity.

Now consider folk biology. Keil (1989) found that preschoolers think that an artifact cannot be turned into a living thing even if its appearance is altered to look just like a living thing, and conversely. But preschoolers do think that appearances can change one
living thing or one artifact into another. Within-domain transformations affect identity, but cross-domain transformations do not. This suggests that preschoolers distinguish artifacts from living things. He speculates that they are innately sensitive to this distinction.

The fact that children think that within-domain transformations can affect identity is hardly surprising. They know about many cases where lexical labels change with appearances: caterpillars become butterflies, boys become men, seeds become trees, construction paper becomes an art project. But children do not experience many cross-domain transformations, and they may be explicitly taught that some of these are impossible. When playing with dolls, for example, parents may say, “That’s not a real baby; it’s make-believe.” Moreover, artifacts and living things are very easy to distinguish perceptually. The living things with which children are familiar have faces, and they tend to be fuzzy, irregular in their movements, and symmetric along a central axis. The fact that these features co-occur may lead to an early, robust, perceptually learned category. By kindergarten, children have been told stories about where entities in this category come from, and they may be reluctant to accept that artifacts can gain admittance by mere change in appearance. I bet their answers would differ if they heard about an animal-like robot coming from Mommy’s tummy.

Finally, consider folk psychology. Kids attribute mental states to others. No one knows exactly how this “mind-reading” ability comes about. It has a relatively fixed time course in development, with the capacity to attribute beliefs emerging after the capacity to attribute desires and perceptions. Mind reading seems to be lacking in apes [Povinelli and Eddy (1996), though see Hare, Care and Tomasello (2001)], and it is selectively impaired in autism [Baron-Cohen (1995)]. These findings indicate an innate domain. Or do they?

Mind-reading abilities may originate in our capacity to imagine things [Gordon (1986), Goldman (1989)]. We can imagine situations that are not actually occurring. This is part of our general executive working memory capacity. Imagining people’s experiences is, at first, like imagining ourselves in another situation. Empathy, emotional contagion, tracking eye gaze, and other relatively primitive, preconceptual abilities may orient us toward others in a way that promotes taking their perspective. If apes lack mind-reading abilities, that may be due to a more general limitation in their capacity to imagine. Apes may have a less developed executive working memory. That would prevent them from actively choosing to imagine what things are like from another perspective.

The sequence of mind-reading development in human children is hardly surprising. Desires and perceptions are usually directed toward perceptually present objects; beliefs are not. Desires and perceptions are associated with characteristic conscious experiences; beliefs are not. Desires and expressions can be attributed using simple accusative verb constructions; belief attributions have sentential clauses as the direct object. All these things make desire and perception easier to learn. Moreover, even if the sequence of learning is fixed, the time course is not. Those who think mind reading is innate emphasize the fact that children begin to attribute beliefs around the age of
four. It turns out that this age varies across cultures. In some cultures, the time course is much slower [Vinden (1999)]. The fact that kids in our culture master belief attribution at four may reflect facts about how children are socialized in the Western world. It may also reflect facts about English and other Western languages. Belief attribution requires embedded clause constructions, which are mastered around the age of four. In fact, good performance on belief attribution tasks is well correlated with the tendency to interpret that-constructions as complement clauses. [Hollebrandse (2003)] In general, degree of social interaction and linguistic skills seem to be the best predictors for mind-reading abilities [Garfield, Peterson and Perry (2001)]. People with autism may be impaired as a consequence of more general social and linguistic impairments.

These remarks are not intended to account for the wealth of findings in support of innate domains. They merely demonstrate that such findings are open to multiple interpretations. Until we have investigated non-nativist explanations of the developmental evidence, nativism cannot be taken for granted. We certainly have many innate capacities, faculties, and biases, but concept-guiding ontological domains may be learned. If they are, then Locke’s empiricist theory of concepts may be closer to the truth than most researchers would dare to imagine.

2.5. Summary

The theory I have been presenting has the following four core tenets:
1. Concepts represent categories via nomological and etiological causation.
2. Concepts are variable constructions in working memory.
3. Concepts are built up of modality-specific memory traces.
4. Concepts, and the core domains that organize them, are all learned.

Tenets 2 and 3 are the central features of Barsalou’s theory. Tenets 3 and 4 are the central features of classical British empiricism. Tenet 1 has been a central theme in contemporary philosophy of mind. None of these tenets enjoys widespread support in psychology. The acceptance of any of them would be a significant departure from the orthodoxy. My goal here has been to show that the orthodoxy may be mistaken. Empiricism is not widely embraced, but it is consistent with current evidence. I would urge researchers to take the empiricist program seriously and test it. If we simply assume that concepts are innate, invariant, and amodal, we may fail to discover important facts about concepts.

If empiricism is a viable theory, why is it so rarely endorsed? One answer is that the majority of people working on concepts in philosophy and psychology believe that there is a fatal objection to empiricism. In particular, many researchers dismiss empiricism on the grounds that it cannot accommodate concepts that represent things that are very abstract. Empiricists claim that concepts are like mental images. That makes sense for concrete categories, but it won’t work for many other cases. Therefore, empiricism is, at best, an incomplete theory of concepts. I will briefly address this concern.
3. The abstract ideas objections

The standard knee-jerk reaction to empiricism is that it cannot handle abstract concepts. Perceptually derived concepts are like mixed media images, and no image, no matter how complex, can depict such lofty abstractions as [TRUTH], [MORALITY], and [DEMOCRACY]. This objection has been discussed in more detail elsewhere [Lakoff (1987), Barsalou (1999), see also Prinz (2002)]. Barsalou, for example, tells a perceptual story about how we understand [TRUTH]. Here, I can only sketch out a response to this objection, illustrating it with [DEMOCRACY] and [MORALITY].

The first thing to note is that non-empiricist theories may have no advantage here. The major difference between empiricism and standard non-empiricist theories is that empiricists say concepts are implemented using modality-specific codes. If concepts were amodal, we would not face the question of how we can depict democracy, but we would face an equally challenging question: How can an arbitrary amodal symbol inside the head represent democracy? In fact, how can it represent anything at all? The appeal to amodal symbols gives the illusion of explaining abstract concepts, but it really makes no progress. We still need to explain how a symbol in the head can represent something complex and unperceivable. If concepts represent by their causal relations to the world, the question becomes, how can a symbol in the head form a causal connection to democracy if democracy is not something one can see, taste, or smell?

Defenders of amodal symbols have a standard strategy for dealing with such cases. They say that abstract concepts are understood by complex networks of inferentially related concepts. We know what to infer if we are told that something is a democracy. We know it is system of governance or collective decision making. We know that, within a democracy, the governed have votes. And so forth.

There are two things to say about this strategy. The first is that it does not require amodal symbols. It just requires symbols. The inferences used to understand what a democracy is can be implemented by a lexical network, relating the English word democracy to other words in English (or some other public language). These words are represented as stored perceptions of sounds or marks or gestures. They are perceptually based, rather than amodal, and the associations between them are learned by listening to discourse about democracy.

This approach to the concept of [DEMOCRACY] suggests that there is a role for labels after all. But they are not necessarily labels in a language of thought. They can be the labels we devise in public languages. Language can be used in this way to expand the power of thinking. Verbal labels serve as placeholders for ideas that are too complex to hold in one’s mind all at once. Labels can also facilitate reasoning, by presenting thought is a logic-friendly, linguistic code. If abstract concepts like [DEMOCRACY] can be adequately explained by networks of labels, then empiricists can explain abstract concepts as readily as their opponents.

The second thing to notice about the label strategy is that it cannot be sufficient on its own. Ultimately, these labels must be pinned down in the senses in order to be applied in the world. If we taught a monolingual Mandarin speaker some sentences containing
the English word *democracy*, she would not thereby know the meaning of the term. To know what the word means, she would have to know where and when to apply it, and she would have to know how to go about making a decision democratically. Mastery of linguistic inferences allows us to reason about democracies, but full mastery of the concept requires some capacity to link it up with the world. The word *vote* may be tied to a stored record of behavioral practices. If someone says, “we are taking a vote” in the right context, we know we should raise our hands. In another context, voting involves going to a voting booth and filling out a form. The sheer variety of practices involved in understanding what a democracy is suggests that we must have a highly variable representation of that category, in accordance with the theory of concepts that I have been defending. Amodal symbols offer no special advantage when explaining these practices. Learning how to vote involve recognizing scenarios and behaving in particular ways. A network of arbitrary symbols, either in a language of thought or in a public language, can contribute something to our competence, but it cannot be the whole story. When it comes to [DEMOCRACY], amodalists have no advantage.

Labels are important for some abstract concepts, but not all. I want to consider a second class of cases: moral concepts [for a detailed discussion, see Prinz (forthcoming)]. Take the concept [MORALLY BAD]. On the face of it, this presents a challenge for empiricists. What does moral badness look like? What image could work? Now, of course, we can all imagine things that are morally bad, and images of certain naughty scenarios may play a role in understanding moral badness. But that cannot be the whole story. We must be able to think of these scenarios that they are bad. What image allows us to do that?

Hume points us toward an answer to this question. He saw that our understanding of morality is intimately tied to certain emotional reactions. Moral badness is related to a variety of bad feelings. This proposal seems to hold up in recent research. When a stranger does something bad, we feel angry, contemptuous, or disgusted. If a person close to us does something bad, we feel disappointed. If we ourselves do something bad, we feel guilty or ashamed. The concept of [MORAL BADNESS] is a disposition to experience one of these emotions in response to actions of ourselves and others. The emotion we experience depends on the nature of the action. Emotions obey a systematic logic. Certain emotions arise in certain circumstances. Rozin et al. (1999) have shown that anger pertains to actions that bring harm to persons or property; contempt pertains to actions against the social order; disgust pertains to actions against nature. To think that something is bad is to be disposed to feel the appropriate emotion and to be disposed to feel badly if you or others fail to feel the appropriate emotion.

To reconcile this account of moral concepts with empiricism, one needs an empiricist theory of what emotions are. I have defended such a theory elsewhere [Prinz (2004)]. The central feature of the account is borrowed from William James (1884). Emotions are perceptions of patterned changes in the body. Emotions get their meaning by being tied to particular kinds of circumstances. Fear represents danger, because it is reliably triggered by dangers. This process need not involve any concepts. A loud noise or a sudden loss of support can trigger fear. Likewise for the “moral emotions.” A glare
can cause anger. But the causes are often more complex. We may get angry at someone who verbally expresses opposition to democracy, for example. The key point is that emotions themselves are perceptual states, and they derive their meaning, like any mental representation, from the kinds of things that cause them to activate. Emotions are easy to accommodate within an empiricist framework, and they extend the range of concepts that we can grasp [see also Barsalou (1999)]. Some of our loftiest concepts may also be our most visceral, and no appeal to amodal symbols can capture the motivational tug and push of moral thinking.

These two strategies, verbal and emotional grounding, do not exhaust the options that are available to the empiricist. Other concepts may be handled in other ways. We need to reflect hard about abstract concepts and investigate each one individually before condemning the empiricist. In each case, I predict, the empiricist will have sufficient resources to handle the concept. Recognizing this is a crucial step in resuscitating empiricism, because most cognitive scientists assume that abstract concepts pose a fatal objection. That assumption is rarely defended, and it may collapse under scrutiny. If it does, conceptual empiricism may regain the centrality that it once enjoyed.

References


Barsalou (this volume)


Ch. 30: The Return of Concept Empiricism

PART 7

GROUNDING, RECOGNITION,
AND REASONING IN CATEGORIZATION
This page intentionally left blank
Chapter 31

CATEGORIZATION, REASONING, AND MEMORY FROM A NEO-LOGICAL POINT OF VIEW

SERGE ROBERT

Université du Québec à Montréal

Contents

Abstract 700
1. Introduction 701
2. Order, Information, and Categories 701
3. Inferences, Arguments, and Information 701
4. Logical and Creative Arguments 703
5. Types of Creative Arguments 703
6. Two Rather Incompatible Views in Cognitive Science 705
7. Experiments on Memory and Logical Competence 706
8. Logical Weakness and Classification of Errors 709
9. A New Theory of Logical Error and Logical Competence 710
10. The Cognitive Functions of Logic 711
11. Corrective Inferences and Cognitive Progress 713
12. The Fundamental Cognitive Function of Logical Reasoning 715
13. The Mind as a Dynamic System: Inference and Memory 715
14. General Conclusions: Categorization, Reasoning, and Memory 716
References 717
Abstract

This chapter discusses the relations between order, information, and categorization. It defines mental operations as inferences on information and makes the distinction between creative and logical inferences. It also shows, based on experiments, that, contrary to what Johnson-Laird holds, logical competence is not dependent on working memory; rather, it depends much more on our capacity to schematize information. Logical error is explained as the introduction of creativity in inappropriate contexts and logic is considered as essential to our knowledge, given its function in the organization and correction of information. The chapter concludes with a conception of the mind as a dynamic cognitive inferential system that adapts to its environment by its capacity to correct its information.
1. Introduction

Much of our knowledge consists of categorization, and reasoning plays an important role in the process of categorization. In this chapter, we will investigate the relations between categorization and reasoning, with some discussion of memory, in order to understand some of the formal mechanisms involved in our capacity to build categories.

2. Order, Information, and Categories

Let us first recall that, according to the laws of thermodynamics, nature has a strong tendency towards disorder – in other words, an entropic tendency. But when a flux of energy passes through a semiclosed medium, an island of order, or negative entropy, appears in the ocean of entropy. However, a medium that is too closed or too open would not get organized. Order can be described as systems, and a system can be defined as distinct entities that are interrelated and that interact in an environment. This is how natural systems such as atoms, molecules, physical objects, living beings, social institutions, and so on are constituted. A good example of nature’s self-organizing capacity is the famous Urey and Miller experiment of 1953, in which amino acids, some of the most basic molecules of life, were spontaneously generated in a flask from a simple constant electrical discharge in a mix of hydrogen (H₂), methane (CH₄), ammonia (NH₃), and water (H₂O).

Information is not natural order in itself, but order as it is perceived or represented. Indeed, some specific natural systems, such as animals and human beings, have the capacity to perceive and/or to represent to themselves that order; thus, they process information. Their actions in their environment are thus mediated by the processing of information about that environment. From this standpoint, animals have a perceptual sensitivity to regularities (or order); they make information out of order; they perceive, represent, transform, store, and communicate information, and incorporate it in actions. In other words, they are sensorimotor systems.

The human mind is a powerful information processor. The processing of information aims at mentally reproducing the structure of natural systems. In trying to represent the organization of systems, it incorporates mental entities, which are categories, mental relations between categories, and mental interactions (or operations) on categories. Thus, categories are very important components of our representation; they are its basic entities, its building blocks. Knowing is categorizing, making and unmaking categories, establishing relations between them, and performing operations on them.

3. Inferences, Arguments, and Information

A mental operation is a transformation that mentally reproduces an interaction between entities. A mental operation can be defined as an inference. To infer is to transform a
piece of information into another piece of information, to go from one category to another. The most basic inferences found in the animal world are associations between perceptions and actions, which are sensorimotor inferences. For example, when the hungry lioness perceives wildebeests, she infers that she should start her hunting strategies: this is a typical sensorimotor inference. More specifically, the lioness has learned to recognize wildebeests, she has hunted them in the past, and she has stored in her long-term memory the association between perceiving a wildebeest and hunting it; consequently, each time she is hungry and sees a wildebeest, she infers that it is appropriate to hunt it. Thus, knowledge requires memorizing and inferring. In our example, one of the inferences involved is the sensorimotor version of the modus ponens logical rule, which goes as follows: if wildebeest, then hunt; there is a wildebeest, therefore hunt. As an association between a perception (seeing a wildebeest) and an action (hunting it), the sensorimotor inference is an association between categories. Categorizing objects and actions, memorizing them, and associating them in inferences characterize knowledge, from its most basic to its most sophisticated forms.

Some complex animals, such as human beings and, up to a certain level, apes, have linguistic capacities, which allow them to produce more abstract inferences that we will call “arguments.” An argument is basically a linguistic inference. It is an abstract mediation between the perception and the action. Since they are linguistic, arguments have components that are propositions. All arguments exhibit the same general structure: from initial propositions, called “premises,” they move on to final propositions, called “conclusions,” by means of a rule of inference. The rule used to realize the inference defines the type of inference, so that the use of a logical rule makes the inference logical, just as a mathematical rule makes the inference mathematical.

The amount of information in an argument and in its propositions can be appraised by the following rule: Information of proposition \( p = 1 - \text{logical probability of } p \). The more probable a proposition \( p \) is, the less information it contains, and vice versa. For example, suppose that I know that “John was in the church at the time of the murder.” We can establish the probability (P) of this proposition (p1) as being \( P(p_1) = 0.5 \) (either John was there or not). Then, the amount of information (I) in p1 is \( I(p_1) = 1 - P(p_1) = 1 - 0.5 = 0.5 \). Suppose now that, progressing in my investigation, I learn that “John was in the church at the time of the murder, praying, in the second row” (p2). The logical probability of this new proposition p2 can be calculated as: \( P(p_2) = 0.5 \) (John may or may not have been in the church) \( \times 0.5 \) (he may or may not have been praying) \( \times 0.1 \) (given that there are ten rows of pews in the church and he could have been in any one of them) = 0.025. Therefore, compared to p1, the information contained in p2 is \( I(p_2) = 1 - 0.025 = 0.975 \). In other words, p2 contains much more information than p1, which is equivalent to saying that p2 is much more improbable than p1, it needs many more conditions to exist if it is to be true. From this perspective, a logical law has a logical probability of 1 and the information it provides is 0: a logical law, or tautology, is absolutely probable but contains no new information.
4. Logical and Creative Arguments

Logical arguments, in which the amount of information does not increase as one moves from their premises to their conclusion, can be opposed to arguments in which the conclusion adds new information to the premises and that we can call, for that reason, “creative arguments.” As Kyburg and Teng (2001, pp. 2–5) clearly state, logical arguments, or deductive arguments, are non-ampliative, truth-preserving, and monotonic. Conversely, creative arguments are ampliative, non-truth-preserving and non-monotonic.

For logical arguments to be non-ampliative means that they do not increase the information contained in their premises. They are also truth-preserving, that is, when their premises are true, their rules of inference are totally reliable, so that the conclusion is necessarily true. The lack of increase in information guarantees their reliability. So, from the truth of “all A are B” and “all B are C,” we can conclude that necessarily “all A are C.” The monotonic character of logical arguments is such that the addition of new information to the premises will not change the truth value of the conclusion. For example, given that from the proposition “P,” we can logically deduce “P or Q” (the disjunction of “P” and any proposition “Q”), no knowledge of any other information that could be added to the premise “P,” such as knowing that “R” or that “S,” would affect the validity of “P or Q,” nor justify any revision of the truth of “P or Q.”

Creative arguments, on the other hand, have the opposite three properties. Being ampliative, they increase our information and this amplification implies that they are not truth-preserving, that they risk adding in the conclusion false information that is not deducible from true premises. This is the case with an induction that would infer, based on the knowledge that all the birds that I have seen can fly, that all birds can fly. The risky component of creative arguments makes them non-monotonic: new information added to the premises could force me to modify the conclusion. Discovering that penguins are birds and that they do not fly and adding this premise to the previous one, would make us conclude, more cautiously than before, that most birds, from which we should exclude penguins, can fly. Bayes’ law in probability theory is a good example of how non-monotonicity can be treated; this law provides a means for measuring the change in the probability of a hypothesis given a new piece of evidence. [Bayes’ law states that $P(H/E) = P(H) \times P(E/H) / P(E)$].

5. Types of Creative Arguments

Influenced by C. S. Peirce, I distinguish three families of creative arguments: analogical, abductive, and inductive. An analogical inference is an intensional generalization, or, in other words, the construction of a general entity based on the resemblances between individual occurrences. According to Holyoak and Thagard (1995, pp. 39–100), different kinds of analogies can be identified: object sameness (O-sameness), relation sameness (R-sameness), and sameness of relation of relations (R²-sameness). In O-sameness, from
the resemblance between different temporal occurrences of similar individual entities, we, can perceive them as successive occurrences of the same object; or from different occurrences of similar individual entities that could be perceived at the same time, we can perceive them as different individuals belonging to the same category. So O-sameness causes us to perceive different occurrences as the same individual object, and different objects as belonging to the same category of objects. We share this sensorimotor capacity with most animals. A lioness’s recognition of an animal as being a member of the wildebeest category is such an O-sameness analogy.

Analogy by R-sameness is an inference that establishes a resemblance between the relationship that A has to B and the relationship that C has to D. This analogy is much more abstract than O-sameness and is specific to animals that are able to use a language, such as apes and ourselves. As Holyoak and Thagard (1995) show, R-sameness creates an abstract unperceived mediation between perception and action that enables its owner to create technologies and to solve problems with a certain ingenuity. Even more abstract are the analogies based on the resemblance of the relation between relations, which seem to be specific to human beings. This R²-sameness is responsible for our capacity to explicitly think causally, to talk of some events as causes and others as effects. All three of these more or less abstract analogies have the same cognitive function: to make intensional generalizations, and to create more or less abstract categories from the perception of similarities.

Tropic figures of speech, such as metaphor and metonymy, are also the result of analogical reasoning, one that uses a source expression to refer to a target reference, given a resemblance in properties between them. This is what happens, for example, when Edmond Rostand, the author of the play *Cyrano de Bergerac*, makes his hero say that because of his cousin Roxane “a dress has passed in my life” [Rostand (1930) p. 241].

Here the source expression *dress* refers to the target *beautiful women*. In this way, tropes create new categories, and more specifically, categories that have an intermediate meaning between the source and the target, as *dress*, for Rostand constitutes a category of beautiful women that have something in them of a woman and something of a dress.

An abductive argument has the following structure: B is the case, and B usually follows A or is accompanied by A; therefore, A might be the cause of B. Thus, we use abductive arguments to develop causal hypotheses between entities. From the frequent co-occurrence or consecutive occurrence of two events, we postulate a causal relationship between them.

An inductive argument is basically an extensional generalization. From the observation of many similar entities, it generalizes their similarity quantitatively to all entities of the same category. So, further to the observation of many white swans, we may conclude that all swans are white. While analogy builds categories, induction fills them up; together, the two kinds of argument contribute substantially to the creation of categories. Like analogies, inductions can be at different levels of abstraction. In our swan example, the induction is sensorimotor, or perceptual, but it can also occur at a more

---

1 The French expression *Une robe a passé dans ma vie* can be translated as ‘a dress has passed in my life.’
abstract linguistic level, or even at the scientific level, when we elaborate a general scientific hypothesis from the observation of some particular cases. As a process of extensional generalization, induction not only contributes with analogy to categorization, but also contributes with abduction to the establishment of hypotheses concerning causal links between categories. In other words, making the inductive hypothesis that all swans are white not only fills the category of white swans, but also makes “swan-ness” the cause of whiteness.

Starting from these distinctions and specifications concerning different types of creative inferences and arguments, we can hold that, in our representation of the world, the basic components are categories and hypothetical causal relations between these categories. Analogy and induction lead us to create categories and, by abduction and induction, we relate these categories causally. Thus, acquiring information about our environment means trying to pin down its structure, which basically involves building categories, putting objects in these categories, and attempting to find causal relationships between these objects and these categories.

6. Two Rather Incompatible Views in Cognitive Science

Now that we have seen the cognitive function of creative arguments, we can ask what the function of logical arguments might be. Creative arguments give us categories and hypothetical causal relationships between them. But given that no new information will arise in the conclusion of a logical argument, what cognitive function could it have? Before answering this question, we need to analyze an incompatibility between two received views in cognitive science: Miller’s (1956) famous thesis on the “magical number” in working memory (or short-term memory) and Johnson-Laird’s (1983) theory of logical error.

One of the most famous theses regarding working memory is of course Miller’s law of seven plus or minus two, which holds that the capacity of our working memory is short, small, and very stable from one person to another, and that it can handle seven plus or minus two elements of information at the same time for about 15 s. On the other hand, Johnson-Laird’s study of human reasoning has led him to posit that, in logical reasoning, we spontaneously make mental models of the premises and use working memory to handle these models in order to reach a conclusion. His thesis is that we make many logical errors and that they result from the weakness of our working memory. So, there are major variations between different individuals’ logical performance and these differences depend on the relative power of our working memory. The more powerful our working memory is, the better we will perform at logical reasoning. The compatibility problem arises because, according to Miller, interindivdual variation (and standard deviations) in working memory are small, whereas, on the other hand, in most experiments on logical competence, by Johnson-Laird and others, interindividual variation and standard deviations are much larger. Johnson-Laird’s research undoubtedly shows that memory plays a crucial role in logical reasoning, but this incompatibility suggests that the process does not work exactly as he presents it.
7. Experiments on Memory and Logical Competence

In order to solve this incompatibility problem, my team performed experiments on working memory and on logical competence in classes of undergraduate university students\(^2\). The experimental sample was made up of 172 undergraduate students enrolled in different programs at my university. The working memory experiments were of a very classical type. First, the students took part in an experiment on visual memory for numbers: for 30 s, they were exposed to a chart on which they could see 15 two-digit numbers; then, they had 1 min to write down the numbers they had memorized. A second experiment was on visual memory for geometrical figures: ten simple figures (circles, triangles, rectangles, etc.) were shown for 30 s and then the students had 1 min to draw the memorized figures. Finally, they were given a test on auditory memory for words: 15 words two or three syllables long were pronounced in front of them, with a 3-s delay between words; at the end, they had to write down the memorized words. The respective means in the results were 6.8 numbers, 7.7 figures, and 8.9 words. Incidentally, the better performance in the third experiment could be explained by a partial use of long-term memory. But, overall, we can say that, as usual, Miller’s magical number was, by and large, confirmed again.

The same groups of students afterwards took part in a logical competence experiment. Before describing this experiment, let us summarize the classical conceptions of logical error. The first theory was that of Woodworth and Sells (1935), which is called “the theory of atmosphere.” It holds that we have a natural tendency in syllogisms to reach a conclusion according to the atmosphere of the premises: thus, we tend to conclude negatively, when at least one of the premises is negative, and to conclude universally, when at least one of the premises is universal. Though this reflex makes many arguments valid, it also happens to make some arguments false. For example, with syllogistic premises of the form “All B are A and all B are C,” we must logically conclude that “some A are C”; however, most of us would conclude erroneously, according to the universal atmosphere, that “all A are C.” Woodworth and Sells’ theory has explanatory power, but it is limited to classical syllogisms and immediate inferences.

A second theory was that of Chapman and Chapman (1959), the theory of false conversion. It relies on the fact that, although universal negative propositions and particular affirmative propositions are simply convertible, universal affirmative propositions and particular negative propositions are, on the contrary, not simply convertible. So, from “no A are B” we can validly go to “no B are A” and from “some A are B” it is valid to go to “some B are A.” But, from “all A are B,” we cannot conclude that “all B are A.”\(^3\) but only that “some B are A,” and from “some A are not B,” no converse conclusion regarding B and A can be logically drawn. According to Chapman and Chapman, our basic logical errors result from a tendency to make a false conversion with universal affirmative or

---

\(^2\) I would like to thank my research assistants G. Choquette, B. Hardy-Vallée, and G. Perreault for their help in performing these experiments. The details of the experiments are being submitted for publication in journals in two articles (Robert et al., submitted a,b).
particular negative propositions. Like Woodworth and Sells’ theory, this theory is cer-
tainly relevant, but, once again, limited to classical syllogisms and immediate inferences.

A third and much more general theory is Johnson-Laird’s [Johnson-Laird (1983)]. It is
also the first procedural theory. According to Johnson-Laird, the procedure that we use in
logical reasoning is to make mental models of the premises. Johnson-Laird’s models are
mainly linguistic and partly schematic and must be exhaustive if the conclusion is to be
valid. What is meant by exhaustivity is a modeling that represents all the information con-
tained in the premises. When exhaustivity requires just a few models, most of us can suc-
cessfully make the inference, but when it needs many models, errors are likely to occur
more often, given that the inference places too much demand on working memory. In these
cases, various models must be eliminated by falsification before we can get to the correct
answer. Such uneasy inferences separate good reasoners from bad ones, which means that
it favors people with a more powerful working memory. For example, “All A are B and all
B are C, therefore all A are C” is an easy syllogism and would be modeled as follows
(parentheses mean that the extension of the term placed in parentheses is exhausted):

All A are B:

(a) b
(b) c

All B are C:

(a) b
(b) c

Therefore, all A are C:

(a) (b) c

On the contrary, “some A are B and no B are C, therefore some A are not C” is a dif-
ficult syllogism. Its modeling, Johnson-Laird style, proceeds as follows:

Some A are B:

a b

No B are C:

(b)
(c)
...;
Therefore, the combined model gives
\[ a (b) \]
\[ a (b) \]
\[ (c) \]
\[ (c) \]
\[ \ldots ; \]

and we try to conclude that no A are C or that no C are A.
However, an alternate model shows that some A might be C:
\[ a (c) \]
\[ \ldots ; \]

Then we try to conclude that some C are not A, but a new model shows us that all C could be A:
\[ a (c) \]
\[ a (c) \]
\[ \ldots ; \]

so we finally reach the only unfalsifiable valid conclusion, some A are not C, by the last possible model:
\[ a (b) \]
\[ (b) (c) \]
\[ \ldots ; \]

Providing logical reasoning with a procedural theory and expressing the procedure as model making are certainly important contributions by Johnson-Laird, but this again raises the issue of the incompatibility discussed above. My attempt to supersede this incompatibility is to try to find an alternative to Johnson-Laird’s models, which appear to be much too sophisticated for the average reasoner. It was in this context that I performed experiments on logical reasoning in a sample of university students, none of whom had any university-level education in logic. They had to answer 19 questions in classical logic, from college-level course of introduction to logic. Some of the questions were about syllogisms, some were problems in the basic logic of classes, and some involved hypothetical reasoning. The subjects were also invited to comment on the procedure they used to arrive at their answers.

As usual in tests of logical reasoning, the subjects’ performance was not outstanding. The success rates were 73.0% in logic of classes, 43.0% in syllogisms, and 58.8% in hypothetical reasoning. A more detailed description of the results in the last category provides some valuable information. There are four basic kinds of hypothetical reasoning, or four ways of asserting the consequences of an implicational first premise. First, there are \textit{modus ponens} and \textit{modus tollens}, which are logically valid inference schemes. \textit{Modus ponens} has the form: “If P then Q, and P, so Q.” \textit{Modus tollens} is a negative version of valid hypothetical reasoning: “If P then Q, and not-Q, so not-P.” The other two
modes of reasoning are invalid; logicians call them “the fallacies of implication.” One is “if P then Q, and Q, so P”; this is the fallacy of the affirmation of the consequent. The other is “if P then Q, and not-P, so not-Q,” which is the fallacy of the negation of the antecedent. For the very easy modus ponens form, our subjects’ success rate was 100%. But for the other three types, it was close to catastrophic: 53.3% for modus tollens, 20.0% for the affirmative fallacy, and only 6.7% for the negative fallacy.

8. Logical Weakness and Classification of Errors

The question that automatically arises when we see such results is why are we so weak at logical reasoning? Is it because we are too sensitive to the atmosphere, or too inclined to make false conversions? Or is it because of the weakness of our working memory? If we use models for reasoning, this will obviously make demands on our working memory. But, given that most of us have poor working memories, why are some of us very weak at logical reasoning and others very strong? In order to answer these questions, we can compare the results of our two experiments and study the correlation between working memory and logical capacities. The Pearson degrees of correlation were as follows: 0.20 between working memory and performance in logic of classes, 0.19 between working memory and categorical syllogisms, and −0.04 for working memory and hypothetical reasoning. Thus, there was no significant correlation between the power of working memory and logical performance in any of the three categories. Indeed, many individuals appeared to be good reasoners and poor memorizers or vice versa.

On the other hand, it is very relevant to analyze the comments the subjects made about the procedures they used to find the answers. A strong correlation between the use of schematic procedures and logical skill is very noticeable. In other words, the subjects who said that they used charts, diagrams, circles, informal kinds of truth tables, and so on were much better at logic. On the other hand, the participants who made fuzzier or more literary comments were much weaker in logical matters. An initial conclusion that can be reached is that the good reasoners are definitely the ones who can translate linguistically formulated premises into figurative representations, rather than the good memorizers.

In an effort to better understand the sources of logical errors, I suggest a classification of their syntactic structure. I hypothesize that we have a tendency to use three specific invalid inferential principles that lead to logical errors, the principles of induction, analogy, and symmetry. The principle of induction (or extensional generalization) can be formulated as “what is valid for some, is valid for all.” For example, from “all the criminals that I have known were liars,” I conclude that “all criminals are liars.” The principle of analogy (or intensional generalization) corresponds to “Objects that are similar are identical.” From “elm trees and beech trees have the same shaped leaves,” we could easily conclude that “elm trees and beech trees belong to the same species.” The principle of symmetry (or abduction) can be stated as “what is logically valid in one direction is valid in the other.” Using this third principle, one could derive the following fallacies: “if you
challenge him, he gets angry; and nobody has challenged him, so he will not get angry” or “if you challenge him, he gets angry; and he is angry, so someone has challenged him.” These latter two fallacies are exactly the fallacies of implication that were explained above: the fallacy of the negation of the antecedent and the fallacy of the affirmation of the consequent, respectively. In this way, the third principle handles an implication, that is, an asymmetric relation between an antecedent and a consequent, as an equivalence – a symmetric implication that holds not only between the antecedent and the consequent, but also reciprocally between the consequent and the antecedent.

9. A New Theory of Logical Error and Logical Competence

Based on our classification of logical errors, we can conclude that the three types of logical errors correspond exactly to the three kinds of creative inferences. I therefore conclude that logical errors result from the use of creative inferences in a context in which we should have applied deductive, monotonic reasoning. We are inclined to make logical errors because we are creative. We make inductive and analogical errors because we can build categories, and we have to build categories in order to know. We make symmetrical errors, like the fallacies of implication, which have exactly the structure of an abduction, because we can be abductive and, in fact, we must be so in order to establish hypotheses of causal relations between our categories.

From my standpoint, the theory of atmosphere applies to a special case of analogical errors, that is, to analogies between the negativity or universality of the premises and that of the conclusion, which works for some classical syllogisms and immediate inferences. The theory of false conversions applies to a special case of inductive errors, and is also appropriate only for certain classical syllogisms and immediate inferences. Johnson-Laird’s theory has a wider scope of applicability and has the advantage of suggesting a procedure, which relies on mental models. Our experiments show that people perform much better on logical tasks, when they use models, but these models must specifically be figurative (or schematic). Johnson-Laird holds that we all have a natural tendency to use mental models in logical reasoning, but his models are much more linguistic than schematic.

My conception is that we have a tendency to make models, but this tendency is not systematic. Many people do not use models and they tend to make errors in logical tasks, since their working memory becomes overloaded, just as Johnson-Laird shows. But others make models that are much more figurative than Johnson-Laird’s, and these people are the best logical reasoners. A figurative model, such as overlapping circles representing interrelated categories, or Cartesian coordinates in algebra, is predominantly visual rather than verbal, so that perception and abstract conceptualization come together. The mental images that constitute figurative models have an intermediate status between perception and language and appear to be necessary for linking language to perception. This idea updates the old Kantian thesis that a schema is a necessary mediation between perception and conceptualization, which allows one to avoid
unorganized perceptions and concepts without empirical content. Equipped with a figurative model of the premises of an argument, we do not need to overload our working memory, and this explains why we can be good deducers without being good memorizers. Compared to such models, Johnson-Laird’s hypothesis appears too linguistic and too artificial, and it places too heavy a burden on our working memory. And this approach probably stems from his symbolist conception of the mind. My perspective, however, is that, as a result of our ontogenetic and phylogenetic evolution, we naturally rely first on our sensorimotor cognitive systems; only slowly, gradually, and with difficulty do we become linguistically and conceptually competent through the unavoidable mediation of mental images. In this line of thought, our cognitive progress in abstraction goes hand in hand with the improvement of our logical capacities, which can be defined as a progressive and unending autonomization of the syntactic structure of argumentation. Thus, all three types of logical errors, which have been classified from a syntactic standpoint, can be seen, from the semantic point of view, as different types of semantic disturbance in the syntactic structure.

10. The Cognitive Functions of Logic

From what we have seen of creative inferences, we can understand that the cumulative use of such inferences builds up systems of categories in our mind. Even though many of these categories can only be juxtaposed or loosely organized, some gradually become logically organized. So, for example, a category with more intension (more properties) and less extension (fewer individual members) can be considered as a subcategory of a category that has a smaller intension and a larger extension; larks, for instance, are a subcategory of birds. Progressively, we learn to organize some categories in logical structures, like Boolean algebras and Boolean lattices. The use of such logical structures, which makes us logically competent, arises from the process of categorization. Logical lattices are structures of order that organize categories by establishing relations between them (such as relations of implication), while logical algebras are structures that perform operations on categories (such as conjunctions or disjunctions). But these logical relations and operations do not create new categories from experience; rather, they organize our categorization of experience into systems.

This brings us to the cognitive function of logical reasoning and back to the question of its relevance, given that it does not increase our information. One of its obvious functions is to make information explicit. Even though logical arguments do not add new information, they make information explicit that was merely implicit in the premises. When, from “all A are B and all B are C” I conclude that “all A are C,” this “all A are C” is not explicitly present in the premises; nevertheless, stating it as a conclusion involves making explicit what was only implicit in the premises. A second and much more important function of logical reasoning is the organization of information: the establishment of relations between categories, so that some become subcategories included in others or supercategories that include others, is typical of such organization.
Organizing categories as elements of a Boolean algebra or of a Boolean lattice is an example of strong structuring between categories. So, logical relations and logical operations organize categories in schematic systems or in networks. This organizational function of logical reasoning implies another function: to supplement memory. We have seen that the use of schematic mental models of the premises of an argument improves our deductive ability. We can now see that this is so because, by making figurative mental models, we structure our information logically and thus compensate for the weakness of our working memory. In the example of a difficult syllogism presented earlier – “some A are B and no B are C, therefore some A are not C” – the valid conclusion can easily be drawn, without overloading working memory, by a simple translation of the premises into overlapping or non-overlapping circles, as follows:

some A are B;

\[ \text{A} \quad \text{B} \]

no B are C;

\[ \text{C} \quad \text{B} \]

So, some A are not C.

This kind of schematization is a prelogical organization and, at the same time, a compression of information. And here is our solution to the incompatible views of Miller and Johnson-Laird: we do indeed have a weak working memory (as Miller claims) and of course it tends to make us bad logical reasoners (as Johnson-Laird points out), but we can compensate for that weakness by a prelogical processing of information – that is, a schematization and compression of information – that compensates for our poor memory and makes us better logical reasoners, better able to avoid the inappropriate uses of the principles of error. The logical processing of categories can apply not only to our working memory, but also to our long-term semantic memory. Indeed, systematically organizing the contents of our “dictionary” (our knowledge of words and their meanings) and our “encyclopedia” (our knowledge of the world) increases our knowledge capacities and compensates for the weakness of our unorganized semantic
memory. The idea here is somewhat as follows: when you understand the structure of a subject matter, you do not have to memorize its different components, because they will come into your mind as a result of deduction, without any effort of memory. A good example concerns someone’s personal library. If a man has a few thousand books in his library, he does not have to memorize the location of each one, which is impossible for his memory; instead, the application of a simple logical system of classification (by discipline, by subdiscipline, by period of time within a subdiscipline, and so on), instead of merely shelving the books at random, allows him to find any of his thousand books in a few seconds. Following this third function of logical reasoning, logic appears as the foundation of our mnemotechnical strategies: a trick for memorizing amounts of information that our memory would otherwise be unable to manage. But this process of information storage by logical organization is probably limited to our working memory and our long-term semantic memory; it might not be at work in our episodic long-term memory, where associations are much more likely to be built by temporal succession or spatial contiguity, or in our procedural long-term memory, in which associations are much more integrated in the body and much more sensorimotor in nature.

11. Corrective Inferences and Cognitive Progress

Given the two families of inferences that we have seen up to now – the creative and the logical – we can explain the acquisition and the structuring of information, respectively. But information also exits from our minds: we forget some information and we also drop some. Of course, the weakness of our memory makes us forget information. But we also deliberately eliminate information, if we judge it to be simply false. When we eliminate information, we need to replace it by better information; thus, elimination of information implies correction of information. This leads me to introduce a third category of inferences: corrective inferences. When we apply our system of categories to new empirical data, we may well encounter counterexamples, which are situations that are inconsistent with our predictions. We then have to correct something in our categorial system. Driven in our representation by three types of creative inferences, we can make three corresponding types of corrective inferences.

We can make an extensional correction (or correction of the induction) when we realize that a previous generalization is excessive: after having held by induction that “all criminals are liars,” one may encounter some criminals who are not liars. We may then pull back and posit that “most criminals are liars.” This kind of correction may at times guide us in the discovery of a causal factor that makes some criminals liars and another factor (or simply the absence of the first factor) that makes other ones not liars. So, extensional corrections can make a causal chain lengthen and branch: from a chain with a cause (being a criminal) and an effect (being a liar), we move to a chain with three links and a fork (from criminal to factor A to liar, or from criminal to factor B to non-liar).

An intensional correction (or correction of the analogy) occurs when a previous analogical association is opposed by newly noticed differences. For example, elm trees and
beech trees, which might have been identified as belonging to the same species, show sufficient differences to be distinguished as two distinct species. In this way, intensional corrections create new categories and new links in causal chains.

Finally, we can make abductive corrections, when we discover the invalidity of an abduction. As we have seen, an abduction goes from “Q and P implies Q” to seeing P as a possible cause of Q. Abductions are implicative fallacies in which an implication is treated as an equivalence. Then P is treated not only as a sufficient condition for Q (the cause of Q), but also as a necessary condition for Q. P and Q are so considered as being in a relation of equivalence. Experience often shows us that this is not the case, that other causal factors may be responsible for the presence of Q; or we may have concluded that the causal relation goes from P to Q and not from Q to P; or that the equivalence between P and Q has to be corrected and be turned into merely an implication between P and Q. Since an abductive correction is the discovery of a new possible cause (other than P) for a phenomenon (Q), it opens up a new causal branch between different causes and a single effect.

To these three correction strategies a fourth one must be added: metacognitive correction. At times, we find our counterexamples to be so disturbing to our system of categories that the addition of a category or of a new branch in their causal relations would be insufficient to integrate the counterexamples into our representation. When that happens, we have the capacity to reorganize the structure of our system of information processing. This happens when we make a logical leap forward, when we move from a simpler logical structure of information processing to a more complex one. This is, for example, what happens when children discover hypothetico-deductive reasoning about counterfactual situations, after being able to argue only on premises they believe to be true. Our capacity to make logical corrections shows that logical competence is progressively acquired through our interactions with our environment and depends on our neuronal plasticity. In this context, our logical competence is an effect of our categorizing activity.

Our information may be true or false, but we do not verify or falsify it by the processes we have just seen above. Nevertheless, it is not totally subjective. First, there is an objective foundation to our creative inferences; moreover, even though we create more information from poor stimuli and may also make logical errors in the processing, we can correct the system. Thus, we can say that correction is what progressively transforms information in objective cognition. In this line of thought, knowledge is neither true nor unfalsified belief; rather, it is corrected belief. Cognitive progress is achievable by successive corrections. From this point of view, cognitive progress results from changes in our system of categorization that increase its explanatory and predictive power. This progress takes the form of a refinement of our system of categories and their causal relations. This category refinement is achieved through the following kinds of corrections: analogical corrections (making new categories), inductive corrections (lengthening causal chains and branching them from causes to effects) and abductive corrections (discovering new possible causes produced by a branching from an effect to many possible causes). In addition to these processes, which build categories and make
their causal relations more complex, we have metacognitive correction, which makes the general formal structures that organize our categories more complex.

12. The Fundamental Cognitive Function of Logical Reasoning

Beyond the cognitive functions of making implicit information explicit, structuring information, and compensating for weakness of memory, logical reasoning is also a condition for making knowledge out of information, because it is a condition of correction. Corrective inferences are required in order to take empirical counterexamples to our hypotheses into account, and these counterexamples appear as failed predictions. Based on a specific organization of our categories, we anticipate experience and we do so by logical deduction. When we believe in a causal relation from P to Q and observe P or not-Q, we hold that if P then Q, and P is the case, so Q should be the case (by *modus ponens*), or Q is not the case, so not-P should be the case (by *modus tollens*). At times, we encounter counterexamples, in the form of inaccurate predictions (observation of not-Q in the *modus ponens* or of P in the *modus tollens*), so that our system of categories requires a correction. But, this correction could not occur without a valid manipulation of the implicative inferences. Indeed, the two fallacies of implication authorize errors in predictions: through these fallacies, we are misled concerning the hypotheses that we test. Fortunately, even young children and bad reasoners use *modus ponens* correctly, even if they have problems with *modus tollens* and the two sophistic schemes. To sum up, logical competence is necessary for testing the truth value of our information; it is a condition of knowledge; our cognitive progress depends on our capacity to correctly manipulate hypothetical arguments. This cognitive function of logical competence is certainly its most important one.

13. The Mind as a Dynamic System: Inference and Memory

The various results obtained so far can be brought together using the theory of dynamic systems. The mind can be thought of as a dynamic cognitive system that contains inferential capacities and memory. Thanks to its memory, it is a reservoir of information, and because of its inferential capacities, this reservoir is dynamic. The mind seizes on informational inputs (our sensations) and rejects informational outputs (memory losses and errors). Creative inferences are input converters that make categories out of sensations. Logical inferences are neutral (non-ampliative) converters, which organize the information inside the reservoir. Corrective inferences and forgetting are output converters.

We know because we forget: otherwise, we would get lost in details. Knowing starts by forgetting most details and using only the most salient aspects of our sensations to construct categories. Half a century of experimental psychology shows us that we have different kinds of memory, and this allows us to distinguish different reservoirs of information: notably, working memory, which can be visualized as a small working
table with very little space (see Miller’s law), and long-term memory, which can be imagined as a large warehouse. In the former reservoir, forgetting happens mainly because of interference, due to the narrow scope; in the latter, erasure by lack of reactivation is more determining than interference. We must add that a logical organization of the information in working memory and in the semantic portion of long-term memory diminishes losses.

From the dynamic systems point of view, cognitive science can be conceived of as a discipline of model making, as it investigates the functioning of the dynamic system of information processing, or of the dynamic system of representation of other natural systems.

14. General Conclusions: Categorization, Reasoning, and Memory

We can now draw some general conclusions. First, we can better understand why computers are sharper than most humans at deduction. It is because they are not creative and do not adapt to the environment. Their system and their programs provide them with preestablished systems of categories. We, on the contrary, have to build our system of categorization step by step to transform it regularly in order to adapt ourselves to our complex and changing environment. And if we introduce creative inferential schemes into computer programs, so that they could transform their categories, then, unlike us, they would not need to use these inferential schemes to adapt systematically to their environment. Our vital need to adapt to the environment is the source of semantic disturbances in reasoning and consequently of logical errors. Adaptation begins with the creation of related categories by creative inferences, which is followed by the organization of these categories, their application to empirical situations, and their correction. The autonomy of the syntactic structure that guarantees logical validity is the object of an endless conquest, because we tend to mix logical contexts of organization and validation of information with creative contexts of discovery of information.

Normative logic, as a formal science of valid inferences, was the instrument for modeling the mind in the symbolist–computationalist conception in the first generation of cognitive science. From such a perspective, the adaptive, dynamic dimension of the mind as a learning machine was neglected. In that school, the capacity to deductively apply a logical structure to empirical data was more studied and more highly valued. In the second generation, in the subsymbolist connectionist conception, our knowledge of lower-level cognitive mechanisms, such as perception, improved, but higher-level components, such as logical reasoning, were neglected. Human cognition is a relatively harmonious mix of both levels and we must move beyond the traditional opposition between these two schools. In other words, contrary to what the computationalist thesis holds, we are not computational devices, and, as the connectionist view shows, we are basically sensorimotor devices, dynamically learning to adapt to our environment. Still, beyond this opposition, we can say that we gradually become relatively computationally competent and that adaptation is precisely the means by which this happens. This reconciliation
standpoint, beyond the traditional opposition, can be called a “neo-logical point of view.” It uses logic as a descriptive as much as a normative tool for the investigation of cognition, so that it can work to explain the progressive construction of our logical competence and of our processing of information as we learn about our environment.

Knowledge consists basically of categorizations and corrections of categorizations so that we can adapt ourselves to our environment. The result of categorization is systems of causally related categories that we store in our long-term memory and transfer to our working memory in our struggle for survival. Logical reasoning is essential to knowledge, because we need it to make implicit information explicit, to organize our information, to compensate for the weakness of our memory, and to correct our information and thereby make knowledge out of it. Categorization and logical reasoning have a common origin in sensorimotor associations between perception and action, which are the basic origin of the process of knowledge. The founding conditions for knowledge are therefore the presence of order in nature and certain marvelous properties of what we call our “minds.” These properties can be summarized as a sensitivity to natural order and some inferential capacities, so that we can produce information about that order; memory, so that we can store the information obtained; and finally, neuronal plasticity, so that we can correct our information and adapt to the mind’s environment.

References

Robert, S., G. Choquette, B. Hardy-Vallée and G. Perreault (submitted a). “Logical competence, working memory and schemata”.
Rostand, E. (1930), Cyrano de Bergerac. (Fasquelle, Paris).
Chapter 32

APPROACHES TO GROUNDING SYMBOLS IN PERCEPTUAL AND SENSORIMOTOR CATEGORIES

ANGELO CANGELOSI

University of Plymouth, UK

Contents

Abstract 719
1. Cognitive symbol grounding 720
   1.1. The symbol grounding problem 720
   1.2. Grounding symbols in cognition 721
2. Linking vision and language: connectionist approaches to category learning and symbol grounding 724
   2.1. Connectionist modeling of category learning and naming 724
   2.2. Connectionist modeling of symbol grounding transfer 727
3. Linking vision, action and language: embodied approaches to language learning and evolution 729
   3.1. Grounding symbols in simulated agents: The symbolic theft hypothesis 729
   3.2. The emergence of language in robots 731
4. Discussion and conclusion 733
References 735

Abstract

This chapter presents the Cognitive Symbol Grounding framework for the grounding of language into perception, cognition and action. This approach is characterized by the hypothesis that symbols are directly grounded in internal categorical representations, while at the same time having logical (e.g., syntactic) relationships with other symbols. The internal categorical representations, constituting the meanings upon which symbols are grounded, include perceptual, sensorimotor, and social categories, as well as internal state representations. Two main modeling approaches to the symbol grounding are presented: (i) the connectionist approach, based on artificial neural networks for category learning and naming tasks, and (ii) the embodied modeling approach, based on adaptive agent simulations and cognitive robots. These models provide an integrative view of the cognitive systems and help our understanding of the relationships between vision, action, and language.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre © 2005 Elsevier Ltd., All rights reserved
1. Cognitive symbol grounding

1.1. The symbol grounding problem

Computational cognitive models that focus on linguistic and symbol-manipulation abilities can use symbols that are either grounded or ungrounded. Models using ungrounded symbols require the interpretation of an external user, such as the researcher, to identify and understand the meaning associated with symbols. For example, in a typical symbolic model of language acquisition, IF-THEN rewriting rules are used to represent the knowledge of the language. In the classical example of past tense learning, rules are used to represent the ability to form the past tense of verbs (e.g., rule 1: “IF regular stem, THEN add suffix –ed”; rule 2: “IF irregular verb to go, THEN use went”). In such models, symbols such as the word go are ungrounded because the interpretative role of the human modeler is required to understand that go means ‘to move or travel.’ Even if the same model had an additional set of symbols that provided definitions of verbs (e.g., a dictionary of verb meanings), these would still be ungrounded, self-referential symbols. This situation is similar to the well-known Chinese room argument [Searle (1980)], where the symbolic task of responding to questions in Chinese without knowing the language can apparently be well performed, even though the subject may not understand the meanings of the questions and answers.

On the other hand, cognitive models based on grounded symbols use words that are inherently significant to the cognitive system. They do not require interpretation by an external user. In such models, the same cognitive agent is able to link the symbols to their own meanings through perceptual, cognitive, and sensorimotor experience. For example, in the grounded model described in Section 2.2, adaptive agents are able to name edible mushrooms, necessary to their survival, with the word eat [Cangelosi and Harnad (2000)]. The same agent is also able to identify the meaning of such a word by referring to the external stimuli, and the corresponding internal states, that produce the word (in this example, the mushrooms that simultaneously possess two perceptual features such as red color and round shape). In addition, the agent’s sensorimotor behavior of approaching and eating the mushrooms also constitutes the grounding of the symbol eat.

The issue of intrinsically linking the symbols used by a cognitive agent to their corresponding meanings has been called “the symbol grounding problem” [Harnad (1990)]. This claims that, for a cognitive model to be considered psychologically plausible, symbols should be intrinsically linked to the agent’s ability to acquire categories through interaction with its environment. In particular, it is essential that at least some basic symbols are directly grounded on sensorimotor categories. Subsequently, new and indirectly grounded symbols (and their corresponding categories) can be formed through the combination of previously grounded basic symbols and categories.

Hybrid symbolic-connectionist models were originally proposed as ideal candidates for solving the symbol grounding problem [Harnad (1990), Sun (2002)]. These would typically include a connectionist module, i.e., a neural network, that deals mainly with the task of grounding the basic symbols in perceptual and categorical representations.
In addition, the hybrid system would use some symbolic modules and information structures (e.g., scripts, lexicon, and episodic memory). Feed-forward neural networks are ideal candidates for performing classification (e.g., categorical) tasks, and the activation of hidden units can be used to analyze their internal, categorical representations. Additional symbolic modules, such as rule sets, can be added on top of the connectionist module to deal with symbol manipulation tasks. These will mostly be used to control high-level cognitive tasks [Miikkulainen (1995)].

More recently, alternative modeling approaches have been proposed to deal with the symbol grounding problem. Some are based on fully connectionist systems. These consist only of a neural network, or a connected ensemble of networks, that perform both the basic grounding and the symbol manipulation tasks. Other approaches are based on embodied agent models and cognitive robotics. They focus on the sensorimotor grounding of symbols. In addition, such embodied models also rely on social learning and interaction between agents (including robots, Internet agents, and humans) for the development of shared grounded symbol communication systems. Some of these robotic approaches also include the use of connectionist networks, while others use different control architectures.

1.2. Grounding symbols in cognition

The various modeling approaches to the symbol grounding problem all have some core features in common. First, each symbol is directly grounded in an internal categorical representation. Internal representations include perceptual categories (e.g., the concepts of red color, square shape, or female face), sensorimotor categories (e.g., the concept/action of grasping, pushing, or pulling), social representations (e.g., individuals, social groups, and relationships), and other categorizations of the organism’s own internal states (e.g., emotional states, motivations). Secondly, these categories are connected to the external world through our perceptual, motor, and cognitive interaction with the environment. Categorical representation of the organism’s internal states can also be mediated by its sensorimotor and cognitive system.

This view of the symbol grounding process will be referred to as “Cognitive Symbol Grounding.” It is consistent with growing theoretical and experimental evidence concerning the strict relationship between symbol manipulation abilities and our perceptual, cognitive, and sensorimotor abilities [e.g., Pecher and Zwaan (in press)]. For example, some authors have explicitly supported the hypothesis that symbols are grounded in our ability to form categories. Harnad (1987, 1990) identifies our innate ability to build discrete and hierarchically ordered representations of the environment (i.e., categories) as the basis of all higher-order cognitive abilities, including language. Categorization of the external and internal world is adaptive to the organisms since it helps them to sort things out and know how to interact with them. This ability is called categorical perception [Harnad (1987)]. In particular, it refers to the process of re-representation of the external environment into internal categories and to the process of “warping” of the similarity space of internal categorical representations. This
re-representational process results in the compression of within-category differences between members of the same category, and the expansion of between-category distances among members of different categories. Categorical perception is a widespread ability in natural and artificial cognitive systems. It has been shown to occur in animals [e.g., Zentall et al. (1986)] and human subjects [e.g., Goldstone (1994)]. The warping effects have also been analyzed in real neural systems [Kosslyn et al. (1989)] and in artificial neural networks [Tijsseling and Harnad (1997), Nakisa and Plunkett (1998), Cangelosi, Greco and Harnad (2000)].

The phenomena of within-category compression and between-category expansion can be graphically represented through the process of the formation of clusters of points in the similarity spaces of categories (Figure 1). Before category learning (Figure 1, left panel), category members produce an undifferentiated similarity space. For example, points representing square objects overlap with those representing circles. The other diagram (Figure 1, right panel) represents the formation of two distinct clusters (cluster of squares vs. cluster of circles) after category learning has occurred. The diagrams represent an abstract two-dimensional similarity space, where each dimension may correspond to some classification component (e.g., geometrical feature) or to the hidden unit activation of a neural network. Relative distances in the similarity space can be calculated using Euclidean measures between points. The two dotted circles in each diagram represent the within-category distances, corresponding to the standard deviation of the Euclidean distances between each point and the center of its cluster. The continuous straight line represents the between-category distance, that is, the Euclidean distance between the centers of the two clusters.

Other researchers have highlighted the relationship between perception, language and action. Barsalou [1999; see also Joyce et al. (2003) for a related connectionist model] supports a view of our cognitive system based on perceptual symbol systems. Perceptual experience, through association areas in the brain, captures bottom-up patterns of activation in sensorimotor areas. In addition, in a top-down manner, association areas partially re-activate sensorimotor areas to implement perceptual symbols. This generates
a mental simulator that produces limitless simulations of schematic representations of perceptual components. Simulators implement a basic conceptual system that supports categorization, produces categorical inferences, and supports productivity, propositions, and abstract concepts.

Coventry and Garrod (2004) propose a cognitive system grounded in both perceptual and action abilities. They hypothesize the online activation of situation-specific models for tasks involving spatial cognition and spatial language judgments (e.g., when subjects are asked to evaluate the use of specific spatial terms). For example, they have extensively studied the appropriateness of the locative prepositions over and above for describing a visual scene depicting a man holding an umbrella in the pouring rain. Experimental and modeling evidence [e.g., Coventry, Prat-Sala and Richards (2001), Cangelosi et al. (in press)] shows that subjects take into consideration a series of factors activated by their previous experience and by the input stimuli involved in the spatial cognition task. These factors include geometric information (relative orientation of the umbrella with respect to the direction of the rain and the position of the human being protected), object-specific knowledge (e.g., typical rain-protection function performed by an umbrella), sensorimotor experience with the objects involved (e.g., force dynamics factors on the direction of the rain).

The grounding of language in action has been extensively studied by Glenberg and collaborators. They have developed an embodied theory of cognition [see also Clark (1997)], where meaning consists of the set of actions that are a function of the physical situation, how our bodies work, and our experiences [Glenberg and Kaschak (2002), Borghi, Glenberg and Kaschak (in press)]. For example, Glenberg demonstrated how language comprehension takes advantage of our knowledge of how actions can be combined, and how linguistic structures coordinate with action-based knowledge to result in language comprehension.

In addition to experimental evidence, the computational approaches to the symbol grounding problem have also provided further evidence in support of the cognitive symbol grounding framework. Various connectionist, robotic, and hybrid symbolic-connectionist models provide a working framework for the implementation of symbol grounding in artificial cognitive systems. The modeling approaches based on classical connectionist networks primarily focus on the grounding in perception and the linking of vision and language. The embodied approaches, based on robots and hybrid robotic/connectionist models, tend to take into consideration both perceptual and sensorimotor components and focus on the link between vision, action, and language. In the next sections, we will review some of these models and will highlight the main findings supporting the cognitive symbol grounding view. The review will mainly focus on models developed by the author and his collaborators at the Adaptive Behaviour & Cognition Research Group\(^1\) of the University of Plymouth (UK). However, other relevant models and simulations will also be briefly referred to and discussed.

\(^1\) http://www.tech.plym.ac.uk/soc/research/ABC
2. Linking vision and language: connectionist approaches to category learning and symbol grounding

2.1. Connectionist modeling of category learning and naming

The great majority of neural network models of symbol grounding concern (i) the use of a dual-route architecture and (ii) the simulation of language production (naming) and comprehension tasks. The dual-route architecture (Figure 2) typically involves both visual input (e.g., retina projection or visual feature list) and linguistic input (e.g., localist or graphemic/phonetic encoding of symbols). The output layer will have symbolic units for representing names and words (e.g., with a phonetic encoding of the lexical items), and a categorical representation of input stimuli (e.g., a localist node for each category, or a visual representation of category prototypes). Some models may use an action-based representation of categories, although this is more typical of connectionist modules within embodied agent systems (see Section 3.1). All input and output layers are connected via some hidden units. The route from visual input to symbolic output is used for language production tasks, such as naming of the objects represented in the visual scene and of their categories. This is the essential route in a symbol grounding network, since the vision→language link constitutes the core mechanism in perceptual grounding. The route from linguistic input to visual/categorical output is used for language comprehension tasks.

One of the first and most influential models of naming and symbol grounding was that developed by Plunkett and collaborators [Plunkett et al. (1992), Plunkett and Sinha

![Typical dual-route architecture for connectionist model of symbol grounding.](image)
The neural architecture was based on the standard dual-route network. The model had two distinct sensory modalities (retinal and verbal) in the input and output layers, and two hidden layers. An autoassociative learning task was used to map both input modalities onto the corresponding output nodes. During testing, either the verbal or retinal input was presented, and the network was asked to produce the output corresponding to the opposite modality. The most interesting result was that training performance was not linearly related to the extent of training. The networks went through stages of sudden improvement, exhibiting something like a “vocabulary spurt” without any apparent reason. This happened both for comprehension and for production, but at different times. The simulation exactly reflects what is observed in children, or in adults when learning a new language: comprehension precedes production. At some stage, the network was able to “understand” which image a name referred to, but not yet to produce this name when shown the corresponding image. But at a later stage, a sudden new improvement in production would be observed.

The connectionist models developed by Harnad, Hanson and Lubin (1991, 1995) specifically focused on the symbol grounding and categorical perception effects. They trained three-layer feed-forward networks to sort lines into categories according to their length. Such lines were represented by input units using two basic coding schemes, iconic (e.g., a length-4 line could be coded as “11110000”) vs. positional (e.g., the same line might be coded as “00010000”). Single bit values could also be more or less discrete (e.g., coarse representations such as 0.1 for 0, or 0.9 for 1 were used). Training consisted of two sequential back-propagation learning tasks: autoassociation and category learning. The first task allowed networks to “discriminate” between different stimuli using a precategorization task with autoassociative learning. The hidden unit activation vectors were examined to record the baseline categorical perception distances for each pair of input patterns. After the autoassociation task, the networks were trained to categorize stimuli by sorting lines into three categories: short, middle, and long. The comparison of baseline and postcategorization distances in the networks’ hidden activations showed the natural side-effect of within-category compression and between-category expansion revealed by human categorical perception. Another point of interest arising from this simulation is that a close scrutiny of hidden representations permits a better understanding of the factors influencing categorical perception. Harnad et al. (1991) found that the distances between hidden unit activations are already maximized during autoassociation, because of the baseline discrimination. This separation, however, is not always so clear-cut as to allow linear separability in the hyperspace of hidden activation, as happens with perfectly categorized stimuli. The back-propagation algorithm, which simulates category learning through supervised feedback, has the effect of “pushing” such unclear representations and forms a hyperplane that separates members of different categories. This results in an improved organization of categorical representations. Tijsseling and Harnad (1997) later replicated these results and also manipulated other factors such as the similarity between stimuli. When there was either extreme nonseparability or extreme separability, no categorical perception effects were present. In the extreme nonseparability case, this is because the task is too difficult to
learn, even at the discrimination level. In the case of extreme separability, compression effects are not present because the task is too easy. In fact, there is no need for category learning because categories already implicitly exist in input stimuli that are extremely separable.

The models described above mostly refer to “naming” tasks, that is, the acquisition of a series of labels for naming (visual) objects and their categories. They generally investigate how representations of name-object associations are learned and how they contribute to the grounding of symbols. However, in human language words are not always directly associated with their referents. In fact, words are mostly associated with other words through syntactic rules and relationships. As shown in (ungrounded) connectionist networks [Elman (1990)] and in studies based on lexical analyses [Landauer and Dumais (1997)], words contain “latent” semantic information that is implicitly expressed through symbol–symbol grammatical relationships. As a matter of fact, the correct definition of symbol implies the presence of symbol–symbol relationships [Harnad (1990), Deacon (1997), Cangelosi, Greco and Harnad (2002)]. These can be logical Boolean relationships, or typical grammatical rules, as in human languages.

Some models of language acquisition have employed a grounded approach that also considers the use of symbol–symbol relationships [e.g., Dyer (1994)]. In particular, various connectionist models have focused on the perceptual grounding of spatial quantifiers. This is because spatial terms have been shown to depend on a variety of semantic and contextual factors such as geometric constraints and extrageometric information [Coventry and Garrod (2004)]. Regier (1996) developed a computational model of spatial prepositions using a method called constrained connectionism. The model learns various spatial prepositions for static (e.g., over and above) and moving (e.g., through) objects, and makes explicit use of the processing of geometrical information. An image of two objects (ground and figure) serves as input to the lower layer of the network. Then the image goes through several levels of geometrical processing. The output units, corresponding to spatial prepositions, are activated according to the geometrical position of the figure object with respect to the central ground.

Cangelosi, Coventry and collaborators [Cangelosi et al. (in press)] have recently developed a connectionist model that produces linguistic descriptions of dynamic scenes involving spatial relations between objects. In addition to the geometric constraints considered in Regier’s model, this new study focuses on the role of extrageometric factors, such as object knowledge and interaction. The model processes 60-s movies showing an object (e.g., teapot) in a specific location pouring a liquid (e.g., water) into a reference object (e.g., cup). The network’s task is to name the objects, and more importantly, to select the most appropriate spatial preposition (e.g., over, above, under, and below) describing the spatial relationships between objects. The model consists of three modules: (i) a neurally inspired vision module based on Ullman-type routines [Ullman (1996)], (ii) an Elman recurrent neural network to learn compressed neural representations of the dynamics of a scene, and (iii) a dual-route network for expressing the names of objects and the spatial terms. The dual-route network is the core component of the model because it integrates visual and linguistic knowledge to
produce a description of the visual scene. The activation values of the linguistic output nodes correspond to rating values given by subjects for the rating of the four prepositions. The multilayer perceptron is trained via error back-propagation, using rating data collected during experiments. Some of the ratings are also used for the generalization test. Simulation results consistently show that the networks produce rating values similar to those produced by human experimental subjects. It also accurately predicts new experimental data on the ratings of scenes, where only the initial frames are shown and the subjects must “mentally replay” the scene and predict its end frame (i.e., where the liquid ends). This model is also consistent with Barsalou’s (1999) perceptual symbolic systems theoretical framework. Currently, a similar model is being extended to deal with further linguistic terms, such as the vague quantifiers *some*, *few*, and *many*. The working hypothesis is that this grounded connectionist approach will permit the identification of the main mechanisms responsible for quantification judgments and their linguistic expression.

### 2.2. Connectionist modeling of symbol grounding transfer

In the connectionist models of category learning and naming discussed above, the focus is on the *direct* grounding of symbols in perception. However, not all symbols need to be directly grounded in perception. In fact, directly grounded symbols can be combined together, through grammatical rules, to produce definitions of new symbols. For example, we can describe a zebra by means of the definition: “A zebra is an animal with a horse shape and stripes.” If a person has never seen a zebra, but has direct grounding experience of the two words *horse* and *stripes*, she can easily infer the perceptual meaning of *zebra*. This process is called “grounding transfer,” i.e., the grounding of basic words is transferred to that of new symbols and categories acquired via linguistic descriptions. This phenomenon has been studied by Cangelosi and colleagues [Cangelosi et al. (2000), Greco, Riga and Cangelosi (2003)] using neural networks. The model consists of a network that has to categorize abstract images consisting of combinations of three different shapes (square, cross, dots) and three different colors (red, green, blue). A modular dual-route neural network was used, in which the hidden layer was organized into two separate groups. The output units indicating shapes were connected only to the first group of hidden units, while those indicating colors had connections solely to the other units. This modular connectivity forces the functional division of the hidden layer into a group dedicated to categorizing shapes and a group that classifies colors.

This type of model is trained through a series of sequential stages: prototype-sorting, entry-level naming and imitation, higher-level learning, and grounding transfer test (Figure 3). In the prototype-sorting and entry-level naming stages, the neural nets are initially trained to categorize and name the color and shape of objects perceived on the retina. In the entry-level naming and imitation stage, an extra learning cycle is executed. This consists of the use of only the symbolic units in both the input and output layers. During the first two training stages, the networks learn through direct trial and error...
experience supervised by corrective feedback. In the third training phase (higher-level learning), networks acquire new higher-order categories solely through symbolic descriptions. New categories are built by combining grounded names. Each description contains the name of a shape, the name of a color, and the name of an object that is new to the network. The grounding test is performed at the end of training by presenting the visual representation of the new object as input.

The networks were able to categorize the colors and shapes of the training stimuli correctly, with a 100% success rate. During the grounding transfer test, novel retinal stimuli depicting new objects were presented to the networks, in order to find out whether grounding had been “transferred” from directly grounded names to higher-order categories. The rate of correct test responses for all networks was 85%. These results indicate that this model is able both to do basic grounding, and above all, to

Fig. 3. Training stages for the symbol grounding transfer simulation.
transfer this grounding to new concepts and symbols. The latter ability greatly depends on the modular organization of the connections between hidden and linguistic output nodes [Greco et al. (2003)]. Moreover, the study further supports a fully connectionist approach to symbol grounding, since the same network also demonstrates symbol manipulation abilities.

3. Linking vision, action and language: embodied approaches to language learning and evolution

Embodied agent models and cognitive robotic research have recently contributed to the issue of symbol grounding. Many robotic models focus on the role of social learning and interaction between agents as the basis of language [e.g., Steels (1999, 2002), Vogt (2002)]. Other robotic [Marocco, Cangelosi and Nolfi (2003), Cangelosi et al. (2004)] and adaptive agent models [Cangelosi and Harnad (2000)] also focus their attention on the cognitive grounding of symbols. However, what all these approaches have in common is their focus on the role of action and sensorimotor knowledge in the grounding of language. In this section, we will describe some of these models to show their contribution to symbol grounding research, and the link between vision, action, and language.

3.1. Grounding symbols in simulated agents: The symbolic theft hypothesis

Cangelosi and Harnad (2000) employed an adaptive agent model to study the role of symbol grounding and categorical perception in the origins of language. They consider two ways of acquiring categories. In the first method, new categories are acquired through feedback-corrected, trial, and error experience with the environment. This is the “sensorimotor toil” approach. Alternatively, new categories are acquired through language, i.e., through hearsay from linguistic propositions provided by language-speaking adults. This is called the “symbolic theft” strategy. In competition, symbolic theft always outperforms sensorimotor toil because it is more efficient (e.g., only one propositional description of a new category is enough to learn it – as in the case of the zebra example). By contrast, repeated experience is required to learn a category by sensorimotor toil. The significant advantage of symbolic theft has been hypothesized to produce an adaptive benefit for language and may help explain the origins of language [Harnad (1996)].

Cangelosi and Harnad (2000) developed a computational model based on simulated adaptive agents using a “mushroom world” scenario [Cangelosi and Parisi (2004)]. This approach is characterized by the simulation of the online interaction between the agent and its environment in a foraging task. The agent’s behavioral and cognitive experience is determined by its embodiment features (e.g., set of sensors and actuators) and the physical aspects of the environment. In this simulation, agents rely on learning categories of foods to survive. For example, mushrooms with feature A (black spots on their tops) should be eaten; mushrooms with feature B (a dark stalk) should have their
location marked; and mushrooms with both features A and B must be eaten, marked, and returned to. All mushrooms have three irrelevant features (C, D, and E) that the foragers must learn to ignore. When organisms approach a mushroom, they emit a call associated with their functionality (“EAT,” “MARK”). Both the correct action pattern (eat, mark) and the correct call (“EAT,” “MARK”) are learned during the foragers’ lifetime through supervised learning (sensorimotor toil). Under some conditions, the foragers also receive the call of another forager as input. This will be used to simulate theft learning of the return behavior.

Organisms’ behavior is controlled by neural networks, similar to those described in the previous sections. Category learning and naming also resemble the process described in Figure 3 (although in a less sequential way). The main difference consists in the ability to process the sensory information about the closest mushroom in order to activate the output units corresponding to the action that must be performed on the environment. The agents learn to categorize the mushrooms by performing the correct action and learning names.

The population of foragers is also subject to selection and reproduction through a genetic algorithm. During the forager’s lifetime, fitness is computed by assigning points for each time a forager reaches a mushroom and performs the correct action on it. At the end of their life cycles, the best foragers with the highest fitness in each generation are selected and allowed to reproduce by engendering five offspring each. The population of newborns is subject to random mutation of their initial connection weights.

To test the adaptive advantage of symbolic theft versus sensorimotor toil, we compared foragers’ behavior for the two learning conditions. In one simulation, the two strategies were directly compared. In the first 200 generations, all organisms learn through sensorimotor toil to eat mushrooms with feature A and to mark mushrooms with feature B. They also learn the names of the basic categories. They are not yet taught to return to mushrooms, or the name of this behavior. In the following 20 generations, organisms live for longer. In the second part of their life, they are divided into two groups: Toilers and Thieves. Toilers learn to return to AB mushrooms in the same way they learned to eat and mark them: through honest toil. In contrast, Thieves learn to return on the basis of hearing the vocalization of the mushrooms’ names (e.g., “EAT” + “MARK” = “RETURN”). They rely completely on other foragers’ calls to learn to return as they do not receive the feature input. To test the adaptive advantage of Theft versus Toil learning, the foragers’ behavior in the two conditions was compared by counting the number of AB mushrooms that are correctly returned to. Thieves successfully return to more AB mushrooms than Toilers. This means that learning to return from the grounded names “EAT” and “MARK” is more adaptive than learning this behavior through direct toil based on sampling the physical features of the mushrooms.

A more direct way to study the adaptive advantage of Theft over Toil was to see how the foragers fared in direct competition against one another. In the second stage after the generation 200 stage, the 100 foragers were randomly divided into 50 Thieves and 50 Toilers, all of whom had to learn to return. Direct competition only occurred at the end of the life cycle, when the 20 fittest foragers were selected to reproduce. The results
consistently showed that thieves gradually came to outnumber Toilers, so that in less than 10 generations the whole population was made up of Thieves.

All these results support the original hypothesis that a Theft learning strategy, based on language hearsay, is much more adaptive than a Toil strategy. This adaptive advantage may have constituted the basis for the origin of language and its adaptive advantage. In addition, categorical perception analyses on the neural network's hidden activation showed that symbolic theft produced enhanced warping effects. The categorical representations for the category “return” in the Toil vs. Theft organisms were contrasted. Data on the Euclidean distance comparisons showed that categories acquired via theft (i.e., via language) have higher between-category distances and lower within-category distances [Cangelosi and Harnad (2000)]. This suggests that Theft learning not only is more advantageous for survival (Thieves collect more return mushrooms than Toilers), but it also optimizes the internal categorical representation by the categorical perception effects. Language learning is based on categorization, but in return it improves categorical learning.

This effect has consistently been reported in connectionist and experimental models of category learning and naming [e.g., Cangelosi et al. (2000), Lupyan (in press)]. In addition, different word classes produce different levels of enhancement of categorical perception effects. For example, in a related model of the evolution of syntax, agents evolve two different classes of words, namely verbs (names of actions) and nouns (names of objects). Comparison of language-induced categorical effects showed that verbs produce more enhanced warping effects than nouns [Cangelosi and Parisi (2004)]. These enhancement effects have been explained by the sensorimotor component of the grounding of verbs [Cangelosi and Parisi (2004)]. Overall, all these results support the importance of modeling the grounding of language in perceptual and sensorimotor abilities.

3.2. The emergence of language in robots

Cognitive robotics has been recently used to model the emergence of language in groups of robotic agents. This has been demonstrated in groups of autonomous robots [e.g., Vogt (2002), Marocco et al. (2003)] and hybrid groups of robots and humans [Steels (1999), Roy, Hsiao and Mavridis (2003)]. These models give one a deeper understanding of the embodied basis of cognition, and in particular of the grounding of language in action [Glenberg and Kaschak (2002)]. For example, in an evolutionary robotic model of the emergence of language, Marocco et al. (2003) showed that the ability to form categories from direct interaction with the environment constitutes the ground for the subsequent evolution of lexicons based on names of actions and names of objects. In this model, agents use proprioceptive and tactile information to actively explore objects in the environment and build categories and a shared lexicon. The controller of each robotic agent consists of an artificial neural network in which, in addition to proprioceptive sensors, two symbolic neurons receive their input from the other agents. The output layer has motor neurons, which control the actuators of the corresponding joints, and two
additional symbolic output neurons, which encode the signal to be communicated to the other agents. A genetic algorithm is used to evolve the agents’ behavior.

The evolutionary robotics model was used to run a series of experiments on the role of various social and evolutionary variables in the emergence of shared communication. Experimental design variables included the selection of speakers (the parent or all peers in the population) and the evolutionary time in which communication is allowed (in parallel with the evolution of manipulation abilities, or following the pre-evolution of good behavior). The simulation results showed that populations evolve stable shared communication mostly when the parents act as speakers and when signaling is introduced in the second stage. Additional analyses of the evolutionary data and the neural network behavior also supported the findings that: (i) the emergence of signaling benefits the agents and the population directly, in terms of improved behavioral skills and comprehension ability; (ii) there is a benefit in direct communication between parents and children, not only because of kinship mechanisms, but also because parents produce more stable and reliable input signals; (iii) the pre-evolution of good sensorimotor and cognitive abilities (i.e., sensorimotor grounding) permits the establishment of a link between production and comprehension abilities, especially in the early generations when signaling is introduced.

A second robotic model [Cangelosi et al. (2004)] focused on human–robot communication. This study simulated epigenetic robots that observe and execute actions via imitative learning. An artificial language was used to communicate the names of actions and objects. Robots first learned a set of basic actions by mimicking them, while simultaneously learning words corresponding to these actions (direct grounding). Subsequently, they learned higher-level composite behaviors by receiving linguistic descriptions containing the words previously acquired. The agents merged basic actions into composite actions by transferring the neural grounding of the words referring to basic actions to the words indicating the higher-level behavior (cf. grounding transfer in Section 2.2).

During training, the imitator robot learned the basic actions and names for opening and closing their left and right arms (upper arms and elbows), lifting them (shoulders), and moving forward and backward (wheels), together with the corresponding words. After few training epochs, robots received first-level linguistic descriptions of combined actions. This consisted of a new word (for the new higher-order action) and two known words referring to basic actions. For example, the action of grabbing the object in front of the agent was described as: “close_left + close_right = grab.” Grounding was transferred from “close_left” and “close_right” to “grab.” Consequently, when the agent was given the command “grab,” it was able to successfully execute the combined action of pushing its arms towards the object and grabbing it. After few more training epochs, the same robot received second-order descriptions. These consisted of a new word (for a novel higher-order action) and a combination of a basic word and a first-order word. For example, “move_forward + grab = carry” combines the grounding of the actions of grabbing (first order) and moving forward (basic order) and produces the behavior of carrying (second order). This higher-level grounding was also successfully transferred to the new second-order word, enabling the agent to correctly perform the action of carrying upon hearing the word carry. The system learned several of these combined
actions simultaneously. Four-word definitions and grounding transfers of up to three levels have also been achieved. This study provides a further demonstration of the process of grounding, and grounding transfer, of symbols in sensorimotor categories in a context of human–robot interaction.

4. Discussion and conclusion

The main features of the Cognitive Symbol Grounding approach will be summarized in this section, together with the most relevant characteristics and contributions of the connectionist and embodied models. We will then be able to consider the remaining research issues and identify avenues for future research directions into symbol grounding.

The Cognitive Symbol Grounding approach is characterized by the following principles:

● A symbol is directly grounded in an internal categorical representation [Harnad (1990)], and at the same time it has logical (e.g., syntactic) relationships with other symbols.
● The internal categorical representations include perceptual, sensorimotor, and social categories, as well as internal state representations.
● Categories are connected to the external world through our perceptual, motor, and cognitive interaction with the environment.
● This view is consistent with current theoretical and experimental psychology research in the grounding of language and cognition in perceptual abilities and embodiment factors [e.g., Barsalou (1999), Glenberg and Kaschak (2002)].

Among the various cognitive modeling approaches to cognitive modeling, those based on artificial neural networks and on embodied agents provide a theoretical framework consistent with the Cognitive Symbol Grounding framework. In particular, connectionist models based on grounded neural networks are characterized by:

● a primary focus on perceptual grounding;
● categorization ability that produces categorical perception effects;
● the use of dual-route neural architectures that permit the simultaneous simulation of language production (vision→language) and language understanding (language→vision/action) abilities;
● the transfer of grounding from directly grounded symbols to higher-order symbols [e.g., Greco et al. (2003)]. This permits the simultaneous control of basic grounding and symbol combination tasks.

The embodied approach to cognitive symbol grounding is mainly based on adaptive agent and robotic models. Such methodologies are characterized by:

● a primary focus on grounding in action;
● the effects of sensorimotor knowledge in the differentiation of symbol (word) classes;
● consideration of the social and interaction factors in the development of shared lexicons;
● analysis of evolutionary factors underlying language and cognition [e.g., symbolic theft hypothesis; Cangelosi and Harnad (2000)].
The two computational approaches to the grounding of symbols and language apparently differ only in terms of which cognitive and social interaction abilities they focus on. In fact, the embodied approach encompasses most of the characteristics of grounded connectionist modeling. This is particularly true for adaptive agent and robotic models that use neural networks to control the agents’ behavior and cognitive systems. All the embodiment modeling examples reported above use neural controllers to organize the sensorimotor, cognitive, and linguistic behaviors of the foraging agents and of the evolutionary and epigenetic robots.

The integration of connectionist networks and embodied agent models has important theoretical and methodological implications. It provides an integrative view of the cognitive system, in contrast to other cognitive modeling approaches that merely simulate isolated abilities (e.g., past tense connectionist model, which only focuses on morphology). This is because all sensorimotor, cognitive, and linguistic abilities are controlled by the same (connectionist) network, or a modular set of networks. This is particularly important for symbol grounding research, because it provides a means of linking vision, action, and language and is consistent with embodiment views of cognition [Varela, Thompson and Rosch (1991), Clark (1997)].

The choice and use of either of the two modeling approaches also has some methodological implications, in addition to the core characteristics listed above. Connectionist models tend to be applied in studies that use cognitive tasks with clear and predefined symbol-meaning sets. When there is a prestructured stimulus set and a predefined number of categories (and names) to which the individual stimuli belong, a typical connectionist simulation will suffice. The training of a connectionist network on category learning and naming is similar to a laboratory study on categorization where the researcher decides which objects and categories to use in the experiment. The adaptive agent and robotic approach, on the other hand, permits the online formation of categories during the organism’s interaction with the environment. This is essential for the simulation of tasks and environments where the process of stimulus grouping and meaning formation is flexible and not defined a priori. For example, simulations of the emergence of communication allow agents to construct an autonomous categorical representation of the environment, while also developing the names for such categories [Cangelosi and Parisi (2002)].

Another methodological implication for the selection of the most appropriate modeling approach depends on the aims of the research. Studies interested in the neural basis of symbol grounding will use connectionist models, or embodied models using neural controllers [e.g., Cangelosi and Parisi (2004)]. Research on the social origins of shared symbols, however, may only require the use of robots with symbolic architectures [e.g., Vogt (2002)].

The computational models discussed in Sections 2 and 3 provide just a few examples of models of the cognitive symbol grounding hypothesis. More models and experimental studies are still necessary to develop a deeper understanding of the mechanisms of meaning formation and the grounding of language in cognitive and sensorimotor categories. To conclude this review, we highlight some of the most promising research directions for current and future studies. These relate to the study of neural mechanisms and modularity, the scaling-up of meaning and symbol sets, and the development of coordinated computational models and experimental studies.
Considerable progress has been made in our investigation and understanding of the neuroscience of language [Pulvermuller (2003)]. Current knowledge of the neural mechanisms for semantic and syntactic processing can be used to design neural network models of symbol grounding inspired by the real-life workings of neurons. For example, an experiment on the modular organization of the organism’s neural network was performed using an adaptive agent model for the evolution of language. Through the application of synthetic brain-imaging techniques [Cangelosi and Parisi (2004)], it was possible to design a modular network architecture that closely (though only qualitatively) resembled the organization of the human speaking brain. The artificial neural networks consistently showed the same functional differentiation between motor areas, which specialize in verb processing (i.e., action names), and sensory processing areas, which specialize in the processing of nouns (object names).

This study also supports research into the modular organization of neural networks for symbol grounding. The work by Greco et al. (2003; see Section 2.2) demonstrated the importance of separating the hidden-output connections that resulted in the modularization of the hidden layer into groups of units specializing in different category classes. Modularity proved essential for the scaling-up of earlier experiments on symbol grounding [Cangelosi et al. (2000)].

Another important issue for future research regards the scaling-up of the cognitive agent’s meaning set, lexicons, and syntactic rules. Current models generally deal with few categories/symbols [e.g., about 10 in Greco et al. (2003)]. Symbol grounding models worthy of addressing language grounding should include a much larger lexicon and richer syntactic structure.

Finally, future research should look toward a more integrative approach to computational modeling and experimental studies. This will lead to direct comparison of simulation data and experimental data. It will also permit the empirical testing of the predictions arising from simulation models. For example, connectionist models of category learning and naming have indicated that language (and category labels) produces enhanced categorical perception effects [Cangelosi et al. (2000), Lupyan (2000)]. This prediction should be investigated in cognitive psychology experiments to assess the psychological plausibility of such an effect. In more general terms, such coordinated empirical and simulation efforts will test the validity for modeling results on the symbol grounding in perception, cognition, and action.

References


This page intentionally left blank
Chapter 33

EMBODIED CATEGORIZATION*

PIERRE POIRIER

UQAM

BENOIT HARDY-VALLÉE

Institut Jean Nicod, Paris, and Lanci, UQAM

JEAN-FRÉDÉRIC DEPASQUALE

UQAM

Contents

Abstract 740
1. Introduction: Embodied categorization 741
2. Purely reactive categorizers 743
   2.1. The perceptual aliasing problem [Whitehead and Ballard (1991)] 744
   2.2. Type I versus II problems [Clark and Thornton (1997)] 744
3. Reactive categorizers that learn 747
4. Representing categorizers 749
5. Emulating and simulating categorizers 752
   5.1. Emulating categorizers 753
   5.2. Simulating categorizers 755
      5.2.1. Simulation of physical categories (folk physics) 757
      5.2.2. Simulation of functional categories (folk biology and mechanics) 758
      5.2.3. Simulation of intentional categories (folk psychology) 758
6. Analogizing categorizers 759
7. Linguistic categorizers 761
8. Conclusion 761
References 762

* Work on this chapter was supported by funding from the SSHRC and FCAR/FQRSC agencies. The authors wish to thank them.
Abstract

In this chapter, we explore the embodied cognitive science of categorization. Embodied cognitive science is a movement within many if not all of the subdisciplines of cognitive science; it seeks to reintegrate, in a principled and nontrivial way, the fact that cognition is a capacity of agents that have bodies, which are situated in an environment. We study a series of increasingly powerful categorizers, seeking to show that the properties of embodiment can be relevant to explain how cognitive systems perform difficult categorization tasks. We start by studying the categorizing abilities of architecturally simple embodied agents. Our guides in the study of these simple agents are AI, AL (Artificial Life), and robotics. Once we are satisfied that embodiment does matter in these simple systems, we study more and more powerful agents, starting with systems endowed with internal states, then systems with a complex internal architecture that provides them with an internal dynamics. Only when we have convinced ourselves that embodiment still matters in these more complex artificial systems, and learned how it does, do we turn to living systems, switching from AI and robotics to neuroscience and psychology as our guides. Our concern, then, will be to show what these living systems have in common with their artificial counterparts, in an effort to determine how embodiment matters to categorization.
1. Introduction: Embodied categorization

Cognitive scientists in fields such as neuroscience, psychology, Artificial Intelligence (AI) and robotics, Artificial Life (AL), and philosophy have recently put forward a view of the mind variously known as situated or embodied cognition, or as extended or embedded or interactive mind. Each of these names stresses a different aspect of the new perspective, but they all share a commitment to the idea that some, and perhaps all, cognitive capacities essentially depend on the body and its environment, for instance on sensorimotor interaction between brain, body, and environment [Clark (1997)]. In this chapter, we will use the expressions “embodied cognition” and “embodied cognitive science” to refer generically to these models. Put simply, the embodied perspective in cognitive science claims that minds have bodies that are situated in environments. Some might say that that is utterly trivial – no one has ever denied that! So why should the embodied perspective matter in cognitive science? And, in particular, why should it matter to the cognitive science of categorization? Our view is that it matters a lot. We have defended that view on conceptual grounds elsewhere [Hardy-Vallée, Poirier and De Pasquale (unpublished manuscript)] and so we will not revisit that issue here. What we shall do instead is defend the view empirically, as it were, by showing how it matters to the cognitive science of categorization.

Categorization is a fundamental cognitive process. Bruner, Goodnow and Austin (1956) see categorization as a solution to information overflow, and Harnad (this volume) even claims that cognition is categorization. Whatever the case may be, it remains certain that categorization is not an abstract task cooked up in the laboratory by cognitive scientists, but an everyday cognitive operation, even for lower animals:

In the simplest case (…), a concept can be seen as a decision procedure where the perception is categorized and the chosen category then determines a choice of action. Gärdenfors (2000, p. 122)

At the simplest level, categorization merges with the capacity to discriminate and react differentially to stimuli:

Any agent in the real world has to be able to make distinctions between different types of objects, i.e. it must have the competence of categorization. Pfeifer and Scheier (1997, p. 157)

As can readily be seen from these examples, the concept of categorization in cognitive science is extraordinarily general – perhaps too general. Given an object’s description, the categorization task is to find the category to which the object belongs (where the set of categories is predefined and the corresponding classes of objects may or may not be disjunct). Mathematically, a categorization is a function from a set of objects to a set of binary vectors, a function that can be described by a table (when the sets of objects and categories are finite).

However, specifying, in the abstract, as it were, the category to which an object belongs and finding an effective procedure that can identify that category are two different things. As computability theory teaches us, there are infinitely more sets than
procedures capable of deciding whether an object belongs to a set (and this is true even with noncontroversial objects like character strings). Discovering these procedures, that is, discovering which procedure applies to a given case and – increasingly frequently in recent times – constructing a procedure where none exists, is a difficult task for a cognitive system. On these questions, there is a remarkable convergence in cognitive science between the more formal (AI) and the more empirical (psychology, neuroscience) studies of categorization. Whereas early researchers in both fields tended to rely on a logical model of categories as necessary and sufficient conditions, a new vision of categorization has progressively emerged, more statistical in nature, describing categorization as a process whereby an agent relies on distributed and neural-like representation [Rumelhart, McClelland and The PDP Research Group (1986)], prototypes [Rosch (1973)], or even simple exemplar bases [Nosofsky (1984)] to extract the spatiotemporal invariants present in objects and situations.

In this context, we say that a categorization procedure is embodied if its success relies in part on the fact that the organism has a body and/or is situated in an environment; for instance, a categorization procedure might be achieved, or learned, only when the system makes use of sensorimotor coordination. Accordingly, an embodied cognitive science of categorization will constrain its explanations of categorization to make explicit reference to the body or environmental situation. Pfeifer and Scheier maintain that “the problem of categorization in the real world is significantly simplified if it is viewed as one of sensory–motor coordination, rather than one of information processing happening ‘on the input side.’” [Pfeifer and Scheier (1997), p. 157]. One can view this chapter as an attempt to evaluate this claim.

We have observed that discovering categorization procedures is a difficult task for cognitive systems. To determine whether it is necessary to constrain explanations of categorization in cognitive science in order to take bodily and environmental properties into account, as the embodiment perspective suggests, we will try to show that the properties of embodiment can be relevant to explain how cognitive systems manage to achieve difficult categorization tasks. To do this, we adopt the following strategy:

1. We start by studying ontologically and epistemically simple systems with simple categorizing abilities. These systems qualify as ontologically simple because they are endowed with a simple architecture involving only a few relevant properties (some of which are embodied). And they qualify as epistemically simple because they are ontologically simple but also because they are artificial, that is, systems that we build and therefore, presumably, systems that we understand best (building

<table>
<thead>
<tr>
<th>Categories\Objects</th>
<th>Object 1</th>
<th>Object 2</th>
<th>Object 3</th>
<th>Object 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category 1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Category 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Modified from Sebastiani (2002).
a system does not make it epistemically transparent – as all neural net modelers are aware – but it does help. Our guides in the study of these simple systems will therefore be AI, AL, and robotics. The categorizing abilities we study qualify as simple because they do not involve any internal state that we could call a representation of the category but are simply manifested in the discriminating behavioral abilities of the system.

2. Once we have convinced ourselves that embodiment does matter in these simple systems, that is, once we can show that the fact that their architecture includes some embodied property is essential to their simple categorizing abilities, we study artificial systems endowed with internal states, and thus more computational and representational power. Since they are artificial, these systems are still epistemically simple (but not as simple as the previous ones since their architecture is more complex), and we will strive to determine how embodiment matters to these systems, even though they now have internal states that could be construed as (abstract, disembodied) representations of categories.

3. Once we have convinced ourselves that embodiment still matters in these more complex systems, we turn our attention to systems with a more complex internal architecture, one that gives them an internal dynamics, while still striving to monitor how embodiment matters in systems with a nontrivial internal dynamics.

4. Only when we have convinced ourselves that embodiment still matters in these more complex artificial systems, and know how it does, will we turn to living systems, switching from AI and robotics to neuroscience and psychology as our guides. Our concern will then be showing what these living systems have in common with their artificial counterparts, in an effort to again determine how embodiment matters.

2. Purely reactive categorizers

In the intelligent agent community, a purely reactive agent is an agent that responds directly to the environment, rather than reasoning about it, using only the current state of the environment to determine what action to perform [Wooldridge (2002)]. The first class of difficult categorization problems we study are those in which objects from two different categories call for different behavioral reactions but where the agent is only provided with ambiguous or conflicting sensory representations of these objects. How can a purely reactive agent behave appropriately, in the face of such stimuli? Before we answer, let us say a few words about how standard, disembodied, neural networks fare with these kinds of categorization problems.

It is well-known that supervised learning allows neural networks to associate an appropriate response with a stimulus. Stimulus–response pairs are presented to a networks, and its connections are adjusted as a function of the error signal generated by the discrepancy between the presented response (the teacher) and the network’s actual response. Recall that, although the order of presentation of the input–output pairs matters a great deal to the capacity of a system to learn a task, and, if it does learn it, to the
quality of its performance (generalization, robustness, etc.), a typical neural network has no influence on the sequence of input–output pairs it is presented with. The categorization network’s abilities are tested on a new set of stimulus–response pairs, by measuring the total or individual error of the categorization task carried out. Neural network categorizers are subject to at least two kinds of problems.

2.1. The perceptual aliasing problem [Whitehead and Ballard (1991)]

Perceptual systems face a perceptual aliasing problem when categorization cannot be perfectly achieved because objects that call for different reactions provide equivalent inputs to the system (and thus it perceives them as identical). In such cases, we say that the stimuli are ambiguous; in general, standard connectionist networks may find a solution that is satisfactory but that does not cover the extension of the categories.

2.2. Type I versus II problems [Clark and Thornton (1997)]

It is well-known that certain network architectures are unable to learn certain categories. This is the case with the one-layer (where we are counting connection layers) perceptron and categories that are not linearly separable, such as the exclusive-or or XOR [Minsky and Papert (1969)]. Clark and Thornton (1997) showed that certain categorization tasks were particularly difficult even for neural networks with more complex architecture. They distinguish between Type-I and Type-II problems. Type-I problems are characterized by the fact that the spatiotemporal regularities needed to categorize an object or situation are immediately apparent to the network, for example, by averaging the input vectors coming from objects that belong to the categories. In Type-II problems, the regularities are not immediately apparent in the input and thus a preliminary recoding of the input is required. Exclusive-or is a typical example of a Type-II problem. In simple cases, like the two-variable XOR, a Type-II problem can be solved by available supervised learning algorithms. But in more complex cases, the task may be close to impossible even for multilayered perceptrons. In such cases, we say that the stimuli are perceptually conflicting. Perceptually conflicting stimuli are stimuli coming from objects that call for different responses but in which the regularities needed to appropriately categorize the objects are deeply hidden in the perceptual stimuli. Thornton (1997) introduced what he calls the Geometrical Separability Index (GSI) to measure how far a given problem is from a Type-I problem. This measure is rooted in the notion of interstimuli similarity, based on Euclidean distance. Where only two categories are involved, GSI is computed by counting the proportion of stimuli that are in the same category as their nearest neighbor (according to Euclidean distance).

Perceptual aliasing and Type-II problems are independent and thus can occur together or not. Some stimuli are ambiguous but perceptually nonconflicting (since they have a high GSI, they are close to a Type-I problem). Conversely, some stimuli may be perfectly nonambiguous but perceptually conflicting (like XOR). In both cases, disembodied connectionist networks can manage pretty well. What happens, however, when
a disembodied connectionist network is faced with a categorization task where the stimuli are both perceptually conflicting and ambiguous? Normally, such systems fail to find an appropriate answer. It is here that embodiment may be of use.

Nolfi and Parisi (1999) describe a number of experiments in which reactive robots and animats must carry out such difficult categorization tasks. A population of robot neural network controllers was “evolved” by a genetic algorithm. Thus, there is phylogenetic training (evolution), but no ontogenetic training (learning). In the first experiment, the robot is in an arena that contains two different cylinders. The robot’s task is to remain close to one cylinder but move away from the other one. However, the cylinders are designed to generate perceptual ambiguity (perceptual aliasing problem): the only difference between the two stimuli is a colored stripe (black on one cylinder, white on the other), which is only visible from certain angles. Categorization is thus complicated by perceptual ambiguity since the two cylinders may appear identical or different depending on the point in space occupied by the robot. In the second experiment [reported in Scheier and Pfeifer (1995)], a robot must approach large cylinders and move away from small ones. However, small and large cylinders appear similar to its sensory apparatus (infrared sensors), and thus the robot is faced with a Type-II problem: interstimuli perceptual similarity (as measured by GSI) is very large. Finally, in the last experiment described here, an animat “lives” in a closed one-dimensional cellular space, in which each cell is characterized by a number representing the stimuli it provides to the animat’s sensors. The space is divided in two 20-cell halves, where each of the 20 cells is labeled with a number from 1 to 20. The animat can move forward or backward and its task is to find the left side of the environment and remain there. Here, the agent faces a stimulus which is both perceptually ambiguous (since two different cells, one on the left side and one on the right, produce identical stimuli) and perceptually conflicting (since the two halves of the space are represented by identical stimuli, the regularity within the classes is necessarily null).

None of these categorization problems can be adequately solved by a disembodied neural network that is passively subjected to the relevant stimuli; that is, no such network can adequately categorize these various stimuli as representing or not representing the target objects (white-striped cylinders in the first case, large cylinders in the second, and the left side of the cellular world in the third). The case illustrated by the third experiment is the most striking since disembodied networks do not manage to find any solution to the problem, all stimuli being both ambiguous and conflicting. Can they be solved by an agent that can rely on properties related the fact that it has a body that is situated in an environment?

Actually, three different solutions are available to an embodied agent in these three experiments. The first experiment illustrates the simplest of these. Recall that the task was to approach the white-striped cylinder and move away from the black-striped cylinder.

1 It is not clear whether this experiment is only a thought experiment, or perhaps a directly programmed system that was not actually evolved, because the authors do not describe its performance or mention the use of genetic algorithms.
This task can easily be solved by any robot that is able to move about in its environment, in particular, any robot that can circle around the cylinders. By circling around the cylinders, the agent becomes able to perceive them from an angle where the color of the stripe is visible and thus to react appropriately. This first strategy exploits the fundamental embodied property of sensorimotor coordination, that is, the possibility for an agent to act in order to select the stimuli to which it will be exposed. In this case, the best thing the robot can do is act so that it will be exposed to an unambiguous stimulus.

The second experiment illustrates another way to exploit sensorimotor coordination. Recall that the robot had to approach large cylinders and avoid small ones and that, because of its limited sensory apparatus, the robot thereby faces a Type-II problem: the classes of stimuli present little regularity that would allow the agent to discriminate between them. Nevertheless, it took only 40 generations for a genetic algorithm to find a neural network controller allowing the robot to correctly discriminate in almost all cases [“almost always,” Nolfi and Floreano (2000), p. 98]. The better adapted individuals, in 86% of the cases, circled around the cylinders and moved away from them if they were small. “This circling behavior is crucial to accomplish the discrimination between the two types of object given that the sensory patterns that the robot experiences while circling the small objects are significantly different from those that the robot experiences while circling the large objects” [Nolfi and Floreano (2000), pp. 99–100]. The role of sensorimotor coordination is illustrated by the GSI, which starts at 0.5 and gradually increases up to 0.9. This indicates that, as artificial evolution developed their capacity to select stimuli through sensorimotor coordination, the stimuli perceived by the agents became more and more regular. Whenever, for instance, a robot faced a stimulus from a large cylinder, it would be less and less perceptually similar to a stimulus from a small cylinder and vice versa. The reason is that, as artificial evolution shaped the robot’s sensorimotor skills, these stimuli were increasingly likely to be the result of behavior appropriate to the categorization task at hand, which made the task easier. By contrast, disembodied passive networks, which are simply exposed to the stimuli without being able to select them, fail to categorize them correctly in 90% of the cases [Scheier, Pfeifer and Kuniyoshi (1998)].

Finally, can embodied agents solve categorization tasks that are at once perceptually ambiguous (perceptual aliasing problem) and perceptually conflicting (Type-II problems)? The animat described in the third experiment, where each stimulus represented an element of one category as well as an element of the other, manages to solve the task presented to it, behaving as indicated in figure 1.

The animat solved the problem by forming a dynamic system with its environment, a dynamical system in which certain environmental states became cyclical attractors. In the figure above, for instance, we see that the animat placed ("born") on the right side of its closed world (the circle) will move toward the left or the right-hand side, but always in the same direction, until it meets, in the left part of the circle, a cell which will send it on to another cell, which in turn will send it back to the first cell and so on. What we have here is a solution in which sensorimotor coordination does not select unambiguous stimuli (there are no such stimuli) or stimuli whose categories are easier
to distinguish (there are none of these either), but creates behavioral attractors that solve the task. Note the extreme niche-dependence of the solution (change the distribution of cells, and the behavioral solution disappears), which is one aspect of the embodiment thesis (albeit pushed to the extreme here).

Embodied categorization thus receives a confirmation of sorts. There are categorization tasks which can be solved by simple cognitive systems, when the use of sensorimotor coordination is allowed, but which cannot be solved when conceived of as problems of passive, disembodied perception in which the system is neither spatially situated nor strongly coupled to its environment. Perceptual aliasing problems are impossible to solve for a passive, disembodied neuronal system. Type-II problems are extremely difficult, if not impossible, for such networks to solve. Finally, the last problem, which combined both types of problems, also cannot be solved by passive, disembodied networks. However, each one of these problems did have a situated solution. We therefore have an instance of embodied categorization, as we had defined it: a situation in which properties related to the agents’ embodiment are directly relevant to explaining their success as categorizers.

3. Reactive categorizers that learn

We discussed the categorizing abilities of extremely simple systems (purely reactive agents) and saw that they are impressive. But purely reactive agents are cognitively limited in a number of ways [Nolfi and Floreano (2000)] and, given the difficulties faced by purely reactive controllers, we should expect embodied but not purely reactive systems to have an advantage over purely reactive systems in more complex categorization tasks. There are a number of ways for agents to be less reactive to their environment. In the following sections, we explore three: learning, internal states, and prediction. The presence of a learning process or of an internal dynamic in a robot’s neural controller allows the robot to react to identical stimuli in different ways, given its previous experience.
In order to explore the relation between learning and evolution, Nolfi and Parisi (1997) devised a fascinating experiment. In a rectangular arena, a robot has to find a target positioned in a random way. It cannot see the target, but the fitness function rewards it if it finds it. Thus, the robot must explore the environment effectively and, to do this, it must avoid colliding with the walls. However, the color of the walls (black or white) changes with each generation. In a dark (black-wall) environment, the robot’s sensors are only activated by the presence of a wall once it is very close to it, which means that the robot must explore its environment cautiously. In a bright (white-wall) environment, the sensors are activated by the walls at a good distance, which means that the robot can explore its environment more rapidly. To adopt the proper search strategy (slow versus quick), the robot must be able to determine which environment it is currently in. Two controller architectures were tested: (1) a purely reactive agent with a simple architecture that cannot learn, and (2) a more complex architecture made up of two modules, an ontogenetically plastic (trainable) action module whose function is to control the motors based on the states of the sensors, and an ontogenetically rigid teaching module, whose function is to produce the training signal for the supervised training of the motor module (with a standard error-reduction rule).

As was to be expected, the robot that could learn performed better, but the interesting question is: why? In order to discover this, Nolfi and Parisi tested mature individuals\(^2\) in both environments, the one in which they had developed and the other one. The individuals that could learn were more efficient in the environment where they had developed, and the question is how this ontogenetic adaptation happens. There are many reasons for this. First, the input stimuli are qualitatively and quantitatively different in the two environments, both with respect to the most frequently stimulated sensors and in the mean activation level of the sensors, and this is true even before learning takes place. Second, immature learning individuals are less fit than nonlearning individuals even in their own environment. Thus, we can conclude that the weights inherited by the learning robots do not simply implement a general solution to the problem (i.e., they are not adapted to both environments) but that they also implement a capacity to learn. The initial states of neural network weights are known to be crucial in learning (for all neural nets, embodied or otherwise) and it appears that artificial evolution may select an initial set of weights that are especially valuable for learning the relevant categories (in other words, the type of environment they are currently in). Moreover, it was observed that immature robots that could learn manifested a juvenile behavior, which disappeared with learning. This juvenile behavior augmented the impact that the qualitative lighting difference between the two environments had on the sensors, thus making the categories less perceptually ambiguous. This characteristic is due to sensorimotor coordination and its use in learning: “…evolved individuals have a predisposition to select useful learning experiences” [Nolfi and Parisi (1997), p. 182].

Nolfi and Parisi (1997) have thus devised non-purely reactive agents (embodied neural networks) that learn how to distinguish between two conditions (“close to a

---

\(^2\) The purely reactive agents were born mature since they could not learn during their lifetimes. Agents that could learn were mature after their training period was over.
wall”/“far from a wall”) in two different environments (“dark environment”/“bright environment”). These agents can thus, in effect, navigate around their world with one or the other set of categories: “close to/far from a wall in a bright environment” and “close to/far from a wall in a dark environment.” By contrast, the purely reactive agents can distinguish between the two conditions but only, as it were, in a generic, average-lighting environment, which is how they treat both the dark and bright environments. As a result, the non-purely reactive agents perform better than their purely reactive counterparts. And it is their ability to select the set of categories appropriate for the environment they were born in that explains their better performance. The question for us is: did the fact that the non-purely reactive agents were embodied (as were their counterparts) play a role in their capacity to learn these categories, and, consequently, must an explanation of the categorizing abilities of the non-purely reactive agents necessarily mention properties of their embodiment? As we saw, these agents are successful at learning which set of categories is appropriate to the environment they are born in because of their innate (i.e., genetically coded) capacity to select stimuli useful for learning. This capacity is encoded in juvenile behavior, that is, behavior that they manifest after birth but that disappears as learning progresses. We could call this kind of learning “embodied learning.” And because this embodied learning is relevant to explaining how the agents can select the right set of categories for the environment they inhabit, the resulting categorization capacity is also an instance of embodied categorization.

4. Representing categorizers

The categorization capacities of reactive agents, even those that learn, constitute more of a “proof of concept” than an actual theory or explanation of categorization. Reactive agents are a fiction of sorts: even very simple organisms have internal states which may be put to use in a categorization task. In agents given the “luxury” of some internal states, states that, in some sense, represent states of the environment, we will see that categorization is achieved by having the internal states and embodied properties cooperate. There is a growing consensus in the embodied cognitive science community that, while some types of cognitive capacities are indeed nonrepresentational, as proponents of reactive robotics [Brooks (1991), Beer (1995)] have claimed, other types of cognitive capacities are representational, as cognitive science has always claimed, but the representations involved are also situated sensorimotor coordination devices. Clark (1997) characterized the change in embodied cognitive science’s concept of representation as a shift from a descriptive model of representation (representation as encoding) to a directive model (representation as control).

Although reactive cognition can be efficient in some contexts, it is not always optimal [Nolfi (2002)], since it requires a good deal of “cooperation” from the environment. Reactive cognition and capacities are a good bet for organisms that exploit or create relatively stable environmental niches. But there are all kinds of instabilities in the environment and environmental instability opens up tremendous opportunities for
organisms that are able to behave efficiently in that niche. We start by exploring agents with limited representational abilities, that is, agents with simple internal states that covary with environmental states, and then turn our attention, in the following section, to agents with more sophisticated representational abilities.

Floreano and Mondada [1996, reported by Nolfi and Floreano (2000)] evolved a controller with a simulated battery that had a limited capacity (20s), which had to be virtually recharged from time to time. Although the agent’s fitness function did not prescribe the recharging behavior (or even the “homing” behavior it presupposed), the evolved controllers were able to produce this behavior. In the experiment, the robot moved about in an arena (a box), in one corner of which there was a recharging area. Behind the recharging area, there was a tower equipped with lamps, indicating the position of the recharging area. The network’s architecture included a hidden layer whose five neurons were fully interconnected. The fitness function was $V(1-i)$, where $V$ is the speed and $i$ the level of activation of the most activated proximity sensor. The fitness of the robot was evaluated at each step, except when it was recharging its battery, and the returned value was added to the preceding sum of fitness values, for a maximum of 150 steps. To maximize its fitness, the robot therefore had to explore the arena [in order to get a high value for $V(1-i)$ at each evaluation] and go to the recharging area when the battery charge had decreased below a certain level (in order to get as much of the possible 150 evaluations as possible – it was impossible to get the 150 evaluations on a single charge). To do that, the robot had to solve two categorization problems. (1) It had to determine when the battery level should be categorized as “Low Battery” (as opposed to “Battery OK”), thus bringing about the homing/recharging behavior (as opposed to the exploration behavior). And (2) it had to determine which part of the arena should be categorized as “the recharging area” (as opposed to “the rest of the arena”). Recall that none of these categories was given to the robot (the homing/recharging behavior was not given to it either). All it was given was the architecture described above and selective environmental pressures that favored speed ($V$) and avoiding the arena’s walls ($1-i$). Nevertheless, given these, evolution found a solution to the categorization problems, and with each generation the lifespan of the robots grew longer and longer.

Floreano and Mondada used a neuroethological method to evaluate the network’s capacities. They placed the robots at various positions in the arena and observed their behavior while recording various internal and external variables (including the activation of neurons and the robot’s absolute position). As a first test, the behavior of the evolved robots was observed as they moved off from various starting positions. The researchers saw that the robots spent as little time as possible in the recharging area, entering it right before their energy was completely spent (note that in this experiment recharging is instantaneous). It was also observed that the robots could not reach the recharging area from certain positions in the arena unless they were already moving at that position, demonstrating the importance of the robot’s internal neural dynamics and its close relation to the structure of the environment. A second series of tests was done while recording the state of the hidden units. By testing the robots in complete darkness, Floreano and Mondada observed that their normal exploration behavior was
quasi-automatic (requiring no input from the environment). However, as soon as the battery became weaker (one-third of the full level), the robot used the luminosity gradient and its memory of its last trajectories to find the recharging area. When darkness is complete, the robot follows a circular trajectory starting from the center of the arena, seeking information about the light gradient. Planning of homing behavior thus begins as soon as the battery reaches the critical one-third level. It was also observed that that the hidden neurons (v-h0, v-h1, v-h2, v-h3, and v-h4) specialized for different functions. The activation of neuron v-h4 covaries with the state of the battery and “announces” that the critical level has been reached. Moreover, v-h4 seems to play a role in planning the homing behavior. Neurons v-h0 and v-h2 are always active except when the robot is approaching a wall. Neuron v-h1 activates after turning, and v-h3 behaves similarly to v-h4. A test under four relevant conditions – two battery-variables (high/low) and two directions (facing the tower/facing the opposite corner) – showed that in each case, the state of v-h4 also covaries with the robot’s position in the environment: v-h4 therefore behaves like a “place-cell.” The authors note that this is similar to findings concerning representations in rats [Tolman (1948)], and in particular, the place-cell v-h4 is reminiscent of the place-cells postulated to exist in the rat hippocampus [O’Keefe and Nadel (1978)].

The robot has a number of embodied categorization capacities but we focus here on only two. The first is an instance of the bodily nature of cognition thesis. We saw that v-h4 behaves like a place-cell, as it covaries with the robot’s place in the environment. But it also covaries with one of the robot’s bodily states: the amount of energy remaining in its battery. Neuron v-h4 in effect tells the robot, “you’re at point p now and your current battery level is x.” Why is this important? We mentioned that, in order to maximize fitness, the robot remains in the recharging area for as little time as possible and enters it right before the battery dies out. This means that the whole area is mapped not in an absolute, disembodied, and viewed-from-above manner, but in relation to the state of the battery: there are areas far from the recharging area, which can only be explored when the battery is full; there are areas the robot finds itself in when the level of the battery is one-third; there are areas it can explore when its battery is quite low; and then there is the recharging area, which it only enters just before the battery dies out. The way the arena is categorized is completely dependent on a relevant “bodily marker”: the state of its battery. Hence, it is impossible to explain how the robot categorizes its environment without reference to this important bodily property. There is no categorization of the environment that is independent of it.

The second embodied categorization capacity of the agent is a more traditional example of sensorimotor coordination. The robot’s homing behavior depends on its categorization of its own behavior as “moving/not moving in the direction of the recharging area,” which in turn depends on gradient climbing. Gradient climbing (or descent) behavior is a typical instance of sensorimotor coordination: a movement (move forward, turn right/left) is performed in order to determine whether it results in more sensor activity (moving up the light gradient) or less (moving down the light gradient). Note that it makes no sense to describe gradient climbing the other way around (input,
then act). Hence, once again, it is impossible to explain the robot’s categorization capacities without reference to its embodiment.

5. Emulating and simulating categorizers

As impressive as are the representational capacities of v-h4, agents endowed with place-cells and the like have no predictive abilities. Such organisms still only react to their environment and cannot anticipate what environment they will be placed in. Yet, prediction and anticipation are some of the most important cognitive devices that organisms have come up with to deal with environmental instability. Prediction is a fundamental feature of any complex adaptive system. And citing numerous neuroscientists, Ryder and Favorov (2001) consider prediction to be one of the brain’s most important functions.

In this section, we discuss the categorizing capacities of representational systems that are able to predict or anticipate the future state of the environment, and thus to plan ahead, on the basis of models they build of that environment. It is our contention that, in order to manage such cognitive feats, organisms must construct cognitive structures that significantly alter their categorization capacities and that how they do so essentially involves the fact that they are embodied.

Together, prediction and planning enable agents to preselect among a set of possible behaviors [Dennett (1994)]. These agents are anticipatory systems, as Rosen (1985, p. 339) defines them: “An anticipatory system S2 is one which contains a model of a system S1 with which it interacts. This model is a predictive model; its present states provide information about future states of S1.” This definition leaves the reference of S1 open: it could be the body, a part of it, an object, an event, an agent, a social institution, in short, anything that can be modeled. To predict and anticipate, agents must construct models of their environment and thus become another kind of representational system, that is, they must become modeling agents. Modeling agents can, to use Popper’s words (1984), let their hypotheses die in their stead.

In all categorization schemas, categorization encodes stimuli by mapping them to types of reactions. In reactive systems, stimuli are mapped to types of actions and thus these agents manifest their categorizing abilities by their discriminative behaviors. In anticipatory systems, categorization maps stimuli to inner structures that act as inner operators on internal processing. Hence, whereas categorization directly serves action (discriminative behavior) in reactive systems, from an embodied point of view it may underlie the forecasting of action inanticipatory systems; that is, it triggers operations that assist sensorimotor coordination. These operations are not behaviors but they aim to control behavior. We will use the generic term “inference”⁵ to cover the range of all internal representational transformations. We discuss two kinds of inner representational control structures and the

---

⁵ Generally speaking, inference is “a cognitive process in which new information is derived from given information” [Hegarty (2004), p. 280]. Nothing in this characterization suggests that inference is something that can only be done by linguistic beings.
categorization processes they effect: emulators and simulators. Emulators and simulators are both modeling systems, and both can be used in categorization. They model the dynamics of bodily (emulator) or worldly (simulator) systems in a discriminative manner in order to prepare for action, and it is this discriminative modeling that constitutes their categorizing process. The terms “emulation” and “simulation” are becoming widely used in cognitive science. Before we get to the point where we cannot understand each other anymore, we should recall how computer science uses these terms.

**Emulation**: When one system performs in exactly the same way as another […]

**Simulation**: Attempting to predict aspects of the behavior of some system by creating an approximate […] model of it.

Emulators and simulators share a core of common properties:
- they are used to model the temporal dynamics of other systems;
- they map sensorimotor inputs and outputs;
- they are updated using real (external) feedback but also using predicted, virtual feedback;
- they function to predict and plan in order to adapt behavior.

Although they are similar in their overall goals, emulators and simulators differ in how they achieve them:
- emulators take idiothetic (internal) cues as input, while simulators take allothetic (external) cues as input;
- emulators model the dynamics of the body and its parts, while simulators model the dynamics of external (nonbodily) objects, events, properties, etc.;
- emulators can be used in dead reckoning, while simulators can be used in piloting.

On this view, the brain *emulates* the body while it simulates the external world. Motor emulators, for instance, predict the next state of the body and can therefore generate feedback signals before they receive actual feedback from the body. Simulators model possible things, agents, and events and their evolution. In emulation, Rosen’s S1 is the body, while in simulation it is the external world. There may be no strict distinction between emulation and simulation, since they are, after all, two kinds of dynamic sensorimotor systems. Everything that applies to emulation also applies to simulation, except that what is simulated is outside the body. We start by isolating the categorization capacities of emulators.

### 5.1. Emulating categorizers

In their 1970 paper, Conant and Ashby present a theorem stating that “every good regulator of a system must be a model of that system” (p. 89), which is now commonplace in

---


5 Piloting and dead reckoning are two basic kinds of navigation used, e.g., in naval navigation: in piloting, the system’s guidance relies on information about distant objects (the stars, the sun, the coastline, etc.), while, in dead reckoning, guidance relies on internal cues (such as the speedometer, watch, and compass). Dead reckoning should be used in unknown environments (e.g., Columbus on his first trip to America) or known environments that one is unable to see (e.g., in a boat in the fog).
control theory and cybernetics. Since the brain is the body’s regulator, some of its inner models could be models of the body. The idea of a body schema [Buxbaum, Giovannetti and Libon (2000)] goes back to Head and Holmes (1911). It has often been invoked as an explanatory concept [say, for phantom limb phenomena, see Melzack (1990)]. But, as Maravita, Spence and Driver (2003, p. R531) remark, the body schema “should perhaps rather be considered as a label for a set of problems still requiring explanation.”

An interesting finding by Schwoebel, Bornat and Coslett (2002) is that body representation is not holistic but is constituted of distinct representational modules. For instance, a widely investigated part of that body schema is motor representation (or “internal models,” as they are usually labeled). Motor representation is the internal modeling of motor behavior in order to guide action. Motor control can be achieved by proprioception, that is, the integration of idiothetic signals arising from cutaneous, muscular, and joint receptors [Farrer et al. (2003)].

The idea governing this trend in research is that tasks that seem simple at first glance reveal themselves to be quite complex, and that complexity can be solved by internal models, “the neural system that predicts force as a function of a given desired state of the limb” [Shadmehr (2004), p. 544]. Take, for instance, the grasping of an object, a kind of fast, voluntary, goal-directed movement. Proprioceptive feedback takes too much time to be of any use for grasping, since signals have to flow from the sensory surfaces to the brain. While proprioceptive feedback takes at least 200 ms [perhaps even 500 ms, see van der Meulen et al. (1990)], the actual correction occurs within 70 ms of the onset of the grasping movement. This is only possible if the hypothesized internal model receives an efference copy of the output command, predicts the next state of the system (i.e., the relevant proprioceptive inputs), and then uses this prediction to guide the grasping behavior. This prediction is feedback from the inside. The model and the body both receive the same input, but the model’s output is computed to correct the grasping movement, faster and more safely than real-world feedback. Hence, these models are emulators. Actual feedback from the environment can then be used to recalibrate the internal model, if necessary. According to Cruse (2003) and Damasio (1999), these emulators are the basis of all world models. Neuroscientists have only begun to unravel their neural implementation [Berlucchi and Aglioti (1997), Maravita et al. (2003)] and their role in body posture [Morasso et al. (1999)] and in phantom limb phenomena [Melzack (1990)].

Internal models are not static, but dynamic. Some basic body schemas are found in newborns [Gallagher et al. (1998)] and these change over a person’s lifetime. From birth to death, limbs and muscles may grow and get weaker or stronger (a process known as plant drift in control theory). Hence emulators must alter their input–output function over time. They can also alter it in order to adapt to tool manipulation [Maravita et al. (2003)], which is important for the process by which cognition extends out into the environment. Emulators can also be learned, and most of the articles cited on internal models deal with motor learning.

---

6 See the work by Shadmehr and colleagues, for instance, Brashers-Krug, Shadmehr and Bizzi. (1996).
From an embodied perspective, what counts as a category depends on the physical and ecological situation: it is a general methodological maxim that one should not map high-level, linguistically formulated categories onto other agents (natural or artificial) or onto lower-level systems like emulators. As a more primitive, lower-level representational device, an emulator is a connection between perception and action. To be sure, it is less direct than reactive categorization (being mediated by internal modeling), but it is still a lot more direct than conceptual categorization.

The work of Jeannerod (1997) and others on prehension gives us some insight into the categories involved in emulation. It is known that, in prehension, the hand is “pre-shaped” both as a function of the grasped object and of the goal and context of the act. Depending on the goal, the objects, and the context, a different motor schema can be initiated according to which neurons are activated. Grasping neurons, holding neurons, tearing neurons, and manipulation neurons have been identified and constitute, as it were, the vocabulary of motor action [Gentilucci and Rizzolatti (1990)]. Grasping, holding, tearing, and the like, may thus be thought of as the categories of the hand motion emulator. Such categories may be represented by a species of what Millikan (1996) calls “Pushmi-Pullyu representations” or what Clark (1997) calls “action-oriented representations,” which are both descriptive and directive. Fadiga et al.’s (2000) characterization of grasping neurons, for instance, is coherent with the motor emulator conception.

While grasping, holding, tearing, etc., seem to be the basic-level categories for emulators, some subordinate levels have been identified. For instance, in grasping, different grips can be generated by emulators. A grip can either be a “precision” or a “power” grip, depending on whether the object held is a pen or a hammer7. In the former case, the thumb is opposed to other fingers, while in the latter all the fingers are opposed to the palm. Hence the motor mind can categorize and react differentially to contexts where-a-precision-grip-is-needed and contexts where-a-power-grip-is-needed. Motor categories like power grip can also be subdivided: for instance, prehension of a sphere and prehension of a cylinder are differentially coded. The inverse model specifies the grip needed, while the forward model anticipates sensory feedback.

Hand dead reckoning in prehension is based on such motor categories. Hence, to categorize for an emulator means to draw a motor inference about what would and what should be the case, from the musculoskeletal system’s point of view. It is fair to expect that many such categories will be found in the nervous system.

5.2. Simulating categorizers

Emulators are internal models of internal dynamics. They may be some of the most basic forms of embodied representational categorization. The emulating dynamics can be embedded in, or combine with, a more general representational process where what

---

7 Recall that prehension is a function of the object and the goal (and context). One can precision-grip a hammer if the goal of the act is, say, to paint with the hammer’s head. Likewise, one can power-grip a pen if the goal of the act is, say, to break it in two.
is planned/predicted is not inside the body but out in the world. When agents categorize things outside their body, they entertain representations of these external things. From the embodied perspective, these representations are not passive reflections or descriptions of external features (mirrors of nature), but controllers that enable the organism’s behavior towards external features. All representational categorizations of this type are geared toward action; they constitute categorization-for-action-forecasting tools. Hence to categorize, at this level, is to map inputs to forecasting internal models. As we have seen, some internal models (emulators) implement action control by planning and predicting bodily states. In this section, we look at the dynamics of internal models (simulators) that implement action control by planning and predicting worldly states.

At this point in our description of the six kinds of categorizers, we still do not presuppose a mastery of language (but see below). The simulating categorizers described here only rely on nonlinguistic sensorimotor representations. Since simulation is the internal modeling of external features, categorization arises when an external process is mapped to a simulator. We will first specify the nature of simulators, and then analyze their role in categorization.

The logic of simulators is similar to that of emulators: instead of waiting for actual (real) perception to provide the signal needed to adjust behavior, cognitive systems sometimes engage in modeling in order to predict what will be the case. In emulators, the aim is predicting actual proprioception while, in simulators, the aim is predicting exteroception. Hence, in some sense, simulation is “virtual perception,” that is, “perception” of a predicted state of the actually perceived system. Simulators are model-based control devices that allow cognitive systems to engage in a specific kind of navigation: piloting. Recall that while dead reckoning is based on idiothetic models (inverse and forward models of the body), piloting is based on allothetic models. A system that can emulate and simulate has information both about its body and about external features of its environment. Moreover, such a system may possess more information about the external world than what it actually gets from sensors, if it can simulate the temporal evolution of some aspect of its environment. In such a case, the system generates a representation of a possible event, based on incoming stimuli from an actual event, about what will be the case (prediction) or what should be the case (planning). Simulation is the reproduction of sensorimotor experiences [see Barsalou (1999), Barsalou et al. (2003)]. To be reproduced, these experiences must be encoded in long-term memory and reactivated when needed. In simulationist accounts, this encoding is not a linguistic recoding: the features of a dog, for example, are not recoded in language-like format, but remain stored in multimodal format, that is, the very format of the actual experience itself. Because there is filtering in attention (selective attention focuses on some features, e.g., movements, vertices, edges, colors, spatial relations, heat, etc.) and in memorization, only schematic properties are encoded (the motor schema for petting a dog, a schematic dog bark, or the generic form of a dog). Hence the reactivation of sensorimotor experiences is also a (virtual) sensorimotor experience.

According to Hommel et al. (2001), the brain principally codes proximal or distal events, events that are perceived or planned. Simulation theory gives us a theory of predicted events.
except that the experience is not driven by external stimuli, but by internal schemata rehearsal. Simulation is a way to anticipate what the world could be if a given action were undertaken. It should be noted that simulation does not mean conscious manipulation of mental images [see Schwartz and Black (1999)]:

- Representations involved in simulation are schematic models (rough replicates), and not images (perfect replicates).
- Conscious mental images are visual representations, while models are spatial representations where nonvisual properties such as force and density can be represented.
- Conscious mental images are holistic, while simulators are schematic.

Although simulation can produce conscious mental images, most of the processing is unconscious [see Barsalou (1999)]. Simulation is neural representation and, as most of the neuroscientific literature indicates, neural representations tend to be unconscious and multimodal. One of the most secure foundations of simulationist accounts is the finding that perception, action, and imagination involve the same brain resources. There is a “common coding of perception and action” [Prinz (1997), p. 167], or to use Jeannerod’s expression, a representation-execution continuum. Many PET and fMRI studies confirm this common coding:

- The neural networks involved when subjects observe or simulate action overlap to a great extent [Decety and Grezes (1999)].
- The neural networks involved when subjects perform or simulate action overlap to a great extent [Jeannerod (1994)].
- Simulated and performed actions retain the same temporal characteristics [Decety, Jeannerod and Prablanc (1989)].
- Simulated actions mainly differ from actually performed actions in bringing about inhibitory processes [in the inferior frontal cortex, see Deiber et al. (1998)] that suppress motor output [Lotze et al. (1999)].

Cognitive systems endowed with simulation capacities are able to categorize events to be, or events to be actualized, by generating the appropriate simulation:

> If the simulator for a category can produce a satisfactory simulation of a perceived entity, the entity belongs in the category. If the simulator cannot produce a satisfactory simulation, the entity is not a category member.  
> Barsalou (1999, p. 587)

Three general forms of categorization can occur: categorizers can simulate physical features, functional features, or intentional features; simulation can be run from the physical, design, or intentional stance [Dennett (1987)]. (We do not claim here that this tripartite account exhausts the range of possible simulations, but it is a convenient way to present this range. Both theorists and experimental researchers will need to contribute to get a more precise account.)

**5.2.1. Simulation of physical categories (folk physics)**

Physical categorizers, those that can adopt the physical stance, are able to simulate geometrical relationships [Kosslyn (1980), Shepard (1994)]. MetaToto [Stein (1994)], for instance, is a robot that uses sensory input to build a map of its environment (i.e., a
graph of the different places visited) and to consult it either online to guide its behavior, or offline to try out behaviors. The robot’s behavior is thus guided by kinematic simulation of the surrounding physical environment, where the forthcoming perceptions (walls, for example) are forecasted. MetaToto categorizes a wall as a limit because of its simulation abilities, not because it reacts when hitting it, as a reactive system would. Although physical categorization thus seems to be an analog of visual experience, a growing body of evidence shows that physical simulations are analogs of action and visuomotor experience. Schwartz and Black (1999) suggest, from behavioral and imagery studies, that physical imagery builds dynamic models, in the physicist’s sense: that is, models that represent physical force fields (gravity, friction, viscosity, etc. of the environment. This is coherent with studies in motor control [Shadmehr and Mussa-Ivaldi (1994)], which hypothesize that representations of force fields must be learned and simulated in order to accurately navigate the environment.

5.2.2. Simulation of functional categories (folk biology and mechanics)

Functional categorizers, those that can adopt the design stance, are able to simulate functional features of artifacts or animals. Hegarty (2004) suggests that most mechanistic inferences are performed by a spatial transformation of the functional features of an artifact rather than relying on a linguistically formulated theory. It is more convenient to visualize the result of pulling a rope with a complicated arrangement of pulleys than to use linguistic reasoning (of course, as complexity increases, language becomes necessary). Functional features need not be objective and fixed features of artifacts. Situated simulating categorizers may sort things out according to their current goals and needs, and thus may create, ad hoc, the category things to bring out of the house on fire “[see Barsalou (1983)].” Most behavior reading in primates can be accounted for as the simulation of functional features. While it has been long assumed that chimpanzees have a theory of mind [Premack and Woodruff (1978)], it now seems more appropriate to describe their social-cognitive expectations as behavior reading [see Heyes and Huber (2000)]. They do not model others’ mental states, but others’ behavioral programs [Byrne and Russon (1998)]. The categories involved are not mental, but categories of movement sequences and subroutines, which are functional. Because they do not have the concept of an unobservable cause [Povinelli (2000)], they cannot posit hidden mental representations, assessable from the intentional stance.

5.2.3. Simulation of intentional categories (folk psychology)

There is an old dispute in cognitive science between the Simulation Theory (ST) of mind reading and the Theory Theory (TT) of mind reading (for an overview of the debate, see the papers in [Carruthers and Smith (1996)] and in [Davies and Stone (1995)]. In ST, knowledge about others’ mental states is not derived from the generalizations of folk psychological theories, linguistically represented, but by “putting oneself in someone’s shoes.” We do not wish to take a position here in favor of ST: all we need is for some
mind reading to be accounted for by ST. Face-based emotion recognition, for instance [Goldman and Sripada (2005)], relies on emotional simulation abilities instead of generalizations from folk psychological, linguistically represented theories. As in the pulleys and rope case, simulation is a more direct way of gaining information about someone else’s internal states, especially when those states are emotionally loaded. Neurological studies of mirror neurons strongly suggest that we ascribe intentional states to others by simulating their behavior and observing the intentional states that result. In the macaque premotor area F5 [Gallese et al. (1996)], it has been observed that, while some neurons (canonical neurons) discharge during observation of graspable objects, others (mirror neurons) are activated during the monkey’s observation of another individual performing an action. These neurons are not sensitive to mere movement, but to action stricto sensu: goal-directed movement. Fadiga et al. (1995) showed that, similarly, observing actions activates the premotor cortex in humans, while PET studies [Rizzolatti et al. (1996)] have located human mirror neurons. Hence the intentional stance is clearly a predictive strategy, which could (but does not always) make use of categories to which we have access not by deriving them from a theory, but by simulating the internal doxastic and volitional states of others on the basis of their behavior, context, and facial expression. Language can give access to higher order intentionality: an agent represents its own mental states as they mentally represent another agent’s mental states, and so on.

6. Analogizing categorizers

According to the embodied perspective, emulation and simulation give us our basic categories. These categories are, at their core, sensorimotor in nature, and thus a big question remains: how can abstract categories – categories that, at first blush, appear to be deeply disembodied – arise from embodied sensorimotor categories? According to Lakoff and Johnson (1980, 1999), sensorimotor categories form the basis of all of our abstract conceptual categories, through a process of cross-domain mapping they call “metaphor”9. For instance, we regularly use expressions like “to grasp an idea” or “to see what someone means” to indicate that we understand something. According to Lakoff and Johnson, this indicates that we comprehend concepts like “understanding,” an abstract psychological category, through analogical inference. Formally, an analogy is a partial identity [Holyoak and Thagard (1995)]: A and B stand in an analogical relationship if and only if there is at least one property that A and B share10. An analogical inference is a “cut and paste” process: from a cognitive domain (the source), copy the structure of an object in the domain and paste it into another (the target), while replacing every variable from the

---

9 Not to be confused with the figure of speech found, e.g., in poetry or political speeches. Although metaphors in poetry are cross-domain mappings, Lakoff and Johnson claim that the process of metaphor is much more general and mundane.

10 Of course, additional constraints must be added to flesh out the formal definition: similarity, structure (consistency and completeness), and function.
source domain by a variable from the target domain. According to Lakoff and Johnson, sensorimotor categories are the source domain of all our abstract categories.

As these remarks indicate, the ability to draw an analogical inference requires that information be able to flow from one domain to another. Mithen (1996) believes that hominid cognitive evolution started when the domain-general intelligence of our primate forebears was supplemented by a few specialized modules (intuitive psychology, biology, and physics), and was completed when these specialized modules became partly “demodularized,” which allowed knowledge to become “available for use in multiple domains of activity” (p. 64). According to the dual-system theory of reasoning [Stanovich and West (2000), Evans (2003)], the human mind is made up of two types of systems. System 1 is a set of intuitive, fast, distributed, parallel, automatic, domain-specific subsystems, similar to connectionist architectures and shared by both humans and other animals. System 2 is rule-based, slower, central (in working memory), sequential, nonautomatic, and domain-general, similar to von Neumann architectures and properly human. Once System 2 is in place, structures in one domain can be exported to a new domain, a cognitive strategy that can be useful to make abstract knowledge possible, on the basis of more concrete categories. The structures need not be linguistic: in Lakoff and Johnson’s framework, the drawing of analogical inference is mediated by schematic (geometrical or topological) models. These models are simulation categories, abstracted out of sensorimotor experiences and applied to other domains. Take, for instance, the Source-Trajectory-Goal schema [from Lakoff and Johnson (1999)] shown in Figure 2.

In this schema, an object O stands at location B, in between location A, where it came from, and location C, where it is going. Such structured representations can be thought of as second-order or relationship prototypes. They apply equally well, for instance, to a thrown stone or a flying bird. The Source-Trajectory-Goal schema can be topologically transformed in order to predict the evolution of B, according to background knowledge. In the two cases at hand (the stone and the bird), the schema is a physical simulation (the bird, even though it is an animal, is categorized as a moving object). In a cognitively fluid mind, however, the Source-Trajectory-Goal schema could also be applied in another domain, for instance the life of an agent, which has a beginning (A), an end (C), and a location between birth and death (B), where the agent currently is. In thinking about the whole course of life, analogical inference makes the category “the course of life” more concrete by exporting a physical structure (the schema) into the biological domain. In their monograph on analogy, Holyoak and Thagard (1995) showed how analogical inferences are fundamental in decision making, science, politics, philosophy, culture, cognition, and everyday life. Lakoff and Johnson

![Diagram](Fig. 2. The Source-Trajectory-Goal schema [Lakoff and Johnson (1999), p. 33].)
demonstrate how abstract concepts may be generated and understood through cross-domain mappings in which sensorimotor categories are the source domain.

7. Linguistic categorizers

We now come to the type of categories and categorizers that are farthest removed from their basic sensorimotor counterparts. Our purpose in this section will simply be to present a speculative account: the story of how someone who accepts the approach presented above might view lexicalized (linguistic) categories or, as we will call them, concepts. Concepts are first and foremost public objects whose usage is controlled by the linguistic community (hence they are normative). These norms may be more or less strictly enforced depending on context. For instance, the category \([\text{GOLD}]\) may be strictly enforced in a commercial or financial context (that is why Archimedes was asked to find out how to distinguish gold from other substances), while it may be less strictly enforced in more mundane contexts (a “gold watch” is not a watch made of gold but a watch whose bracelet and some other parts have been plated with gold). The conceptual system (that is, the system of concepts) of a community, including their norms of application, may also be refined over time, by trial and error (e.g., mushroom categories), by memetic evolution (perhaps folk psychological categories), or by explicit design, either by blending (e.g., the category \([\text{WIRELESS CABLE}]\)) or by pure invention (\([\text{PROTON}]\)). In some communities, institutions may be created to control, promote, and develop conceptual systems.

When a child is born, she does not possess any concepts (as we define them: lexicalized categories), although she may possess, or soon develop, many of the categories mentioned above. Although she does not possess them, she is however born in a linguistic community made up of competent users of the lexicalized categories. Thus, her environment, the environment she will eventually simulate, not only contains worldly events but also word-category contingencies, that is, contingencies where the lexical item reserved by the community for a given lexicalized category is used in the context of talking about an object (event, etc.) that belongs to that category. By learning to simulate these contingencies, the child will come to possess an internal model of her community’s lexicalized categories and thus will be able to use them as they were intended to be used by her community.

We realize that this presentation of what are, for us language speakers, the most important categories (or, at least, those with which we are most familiar) is altogether too brief. Our purpose was not to discuss the issue in any depth but to find a place for these categories within the system of categories and categorizers we have been discussing.

8. Conclusion

In conclusion, we would like to do some model-based prediction of our own and anticipate the future of the study of categorization within cognitive science. In Hardy-Vallée
et al. (unpublished), we expressed the sentiment that it was time to combine cognitive computational neuroscience and evolutionary-developmental cognitive science. It should be clear from this chapter that we also believe the time has come to add the embodiment perspective to the mix: embodied evolutionary-developmental computational cognitive neuroscience (obviously, no one will ever use this cumbersome descriptor to denote the discipline, but we feel that this is where the field is moving and that, one day, this is what people will have in mind when they think of cognitive science).

Embodied evolutionary-developmental computational cognitive neuroscientists believe that, to fully understand cognition, complete models, that is, models of embodied and situated agents will prove essential. This methodological rule directly follows from the thesis that cognition emerges from the interaction between brain, body, and world. If cognition does emerge from the interaction between these three components, then only models that include all three will ultimately allow us to grasp the nature of cognition. Since it is currently impossible to build models and simulations that are both complete and detailed, simplifications must be made and certain things have to be excluded. Indeed, embodied cognitive scientists believe that a measure of understanding will be gained by studying simple and superficial models of complete agents. In some cases, the models we have reviewed here are indeed extremely simple: the organism may be a point-animat living in a grid world, whose behavior is determined by a one-layer feed-forward perceptron! Nevertheless, simple models like these can help us understand some general principles governing categorization. If anything, simple models show how categorization capacities that are quite sophisticated can emerge from very simple embodied and situated systems. Of course, to have any significant impact on the cognitive science of categorization, embodied cognitive science will need to go beyond simple systems and general principles of categorization. As the field matures (and as computational power grows), more detailed models of the brain will have to be included in models of complete agents, as well as more detailed models of the body, its development and evolution, and the environment. Until we can build these detailed models of complete agents, a coevolution between the detailed but restricted models of psychology and neuroscience and the superficial but complete models of embodied cognitive science (robotics, etc.) is desirable. Such is the incremental strategy adopted by roboticists to build intelligent systems and we hope we have shown how the strategy can be pursued in order to further our understanding of one of the most important cognitive capacities of intelligent agents: their ability to categorize.

References


Harnad, S. This volume.


Chapter 34

CATEGORIZATION OF OBJECTS, SCENES, AND FACES THROUGH TIME

ÉRIC McCabe, CAROLINE BLAIS and FRÉDÉRIC GOSSELIN

Département de psychologie, Université de Montréal

Contents

Abstract 768
1. A model of categorization 769
2. Basic-level literature 772
3. Discrete processing cycles 773
   3.1. A Bubbles primer 773
      3.1.1. Stimulus Set? 773
      3.1.2. Stimulus generation space? 774
      3.1.3. The samples? 774
      3.1.4. The task? 774
      3.1.5. Response? 774
      3.1.6. Observers? 775
   3.2. Fossilized discrete processing cycles 775
   3.3. What can temporal bubbles reveal about a SLIP categorizer? 777
4. The need for flexibility and a paradox 778
   4.1. Limited processing capacity 778
   4.2. The need for flexibility 779
   4.3. Back to the paradox 780
5. Categorization as an iterative process 780
   5.1. Compulsory feedforward processing sweeps 781
      5.1.1. The nature of the information processed during compulsory feedforward sweeps 782
   5.2. Flexible iterative processing sweeps 782
      5.2.1. Empirical evidence for flexible and iterative processing sweeps 784
      5.2.2. Deactivation studies 786
6. General discussion 786
References 788
Abstract

In this chapter we describe the Strategy Length and Internal Practicability (SLIP) model, a formal model for thinking about categorization, particularly categorization through time. We also discuss an early application of this model to basic-levelness. We present novel evidence for discrete processing cycles through time. We then turn to aspects of categorization through time that have been neglected in the categorization literature: our limited processing capacities – the necessity of having a flexible categorization apparatus – and the paradox that this inexorably brings about. Finally, we describe a twofold resolution of this paradox. Throughout, but especially toward the end of this discussion, we attempt to bridge work done in categorization, vision, neuropsychology, and physiology.
1. A model of categorization

Figure 1 shows four artificial scenes synthesized by combining two different luminance patterns (that we call flat and hilly) with two different chromatic patterns (labeled grassy and sandy). This artificial example captures some of the essential characteristics of real-world categorization. These stimuli can be categorized as either “field” (the combination of is_flat and is_grassy), “desert” (is_flat and is_sandy), “mountain” (is_hilly and is_grassy), or “dune” (is_hilly and is_sandy) at the most specific level of categorization. At the most abstract level of categorization there is more than one possibility – we will come back to this point later – but, for now, let us only consider the two categories “flat” and “hilly” (Figure 1, Figure 1 is Plate 34.1 in the Separate Color Plate section). Thus, the “mountain” and the “dune” scenes are also “hilly,” and the “field” and “desert” scenes are also “flat.” We have a small category hierarchy. And we are ready to begin the unpacking of our ideas about categorization through time.

The Strategy Length and Internal Practicability (SLIP) categorizer model applies “optimal” testing strategies to determine the category membership of objects. The goal here is not so much to mimic human performance precisely but to provide a nonarbitrary starting point for future modeling efforts as well as a framework to better understand human performance [e.g., Anderson (1990, 1991), Kersten (1990), Feldman (2000)]. A strategy comprises sets of noisy detectors. For example, the desert scene illustrated in Figure 1 satisfies two category strategies: Strat(“dune”) = [{is_hilly} and {is_sandy}], is the SLIP strategy for the “dune” category and comprises two sets of detectors; Strat(“hilly”) = [{is_hilly}] is the SLIP strategy for the “hilly” category and comprises a single set of detectors. We think of these sets of detectors as populations of specialized neurons (e.g., in V4 for color, in V5 for motion). SLIP launches a subset of all these detectors in parallel. The size of this subset is related to the amount of information that humans can process simultaneously. We discuss this point in some detail in a subsequent section.

Because the detectors in a set are redundant, only one of them needs to be successful to verify the entire set. For example, to verify that a scene is “flat” in the category hierarchy illustrated in Figure 1, one successful luminance detector suffices. Everything else being equal, SLIP predicts that strategies associated with more redundant sets of detectors will have a higher probability of being completed after few discrete processing cycles (t). There are two ways of increasing the redundancy of a strategy: either more detectors of “feature X” become available, or more exemplars of “feature X” become available.

Often, more than one set of redundant detectors is required to place a scene in a category. For example, to verify that a scene is a “dune” in the category hierarchy displayed in Figure 1, one successful luminance detector and one successful chrominance detector are required. Everything else being equal, SLIP predicts that categories associated with shorter strategies will have a higher probability of being completed after few processing cycles.
We now turn to the formalization of these ideas. The cumulative probability that a strategy comprising $n$ sets of redundant detectors is completed at processing cycle $t$ or earlier is given by

$$
\mu = \prod_{j=1}^{n}(1-\phi_j),
$$

(1)

with $\phi_j = 1 - \gamma_j$. The constant $\gamma_j$ is the probability that the set of detectors $j$ is successful after one processing cycle. It is the “weight” given to dimension $j$ [for details see Gosselin and Schyns (2001b)].

Although the probability distribution of Equation (1) is useful in its own right, we have more often employed its associated density function, the function that gives the probability that a SLIP strategy is completed after exactly $t$ processing cycles. To compute it, we must subtract two cumulative probabilities: the probability that the strategy is completed in at most $t$ processing cycles minus the probability that it is completed in at most $t-1$ processing cycles:

$$
\chi = \mu(t) - \mu(t-1).
$$

(2)

To illustrate Equation (2), we have applied it to Strat(“hilly”) and to Strat(“dune”). The predicted density functions are given in Figure 1. Here, we have assumed that both $\phi$’s (i.e., the probability that the set of detectors $j$ is unsuccessful during one processing cycle) are equal to 0.5. Two things are remarkable about these density functions: (1) They are very different from one another. In particular, they differ on the average number of cycles necessary before verification. The mean values of such density functions are equal to $\sum_{t=1}^{\infty}t\chi(t)$, i.e., 2 cycles for Strat(“hilly”) and 2.67 cycles for Strat(“dune”).

---

Fig. 1. (a) The four scenes used in this experiment and the corresponding low-level category names (“field,” “mountain,” “desert,” and “dune”) that were learned by all participants, and the high-level categorizations (“flat” and “hilly”) that were learned by participants given luminance cues (LUMI participants). The two histograms illustrate the RT curves predicted by a SLIP categorizer model. (b) The proportion of the four error types as a function of presentation time for each observer group. The solid and dashed lines are, respectively, the average best-fit curves from individual data points applying the SLIP model to the LUMI participants and to participants given chrominance cues (CHRO participants). The black curves represent the proportion of errors on none of the perceptual dimensions (e.g., respond “field” when presented with a field scene); the green curves represent the proportion of errors on the luminance dimension (e.g., respond “mountain” when presented with a field scene); the red curves represent the proportion of errors on the chrominance dimension (e.g., respond “desert” when presented with a field scene); and the blue curves represent the proportion of errors on all dimensions (e.g., respond “dune” when presented with a field).
Another remarkable aspect of these density functions is their shapes, which are reminiscent of response time density functions [e.g., Luce (1986)]. If we assume that response time \((RT)\) is a linear function of the number of processing cycles (i.e., \(RT = a \times t + b\), where \(a\) and \(b\) are two free parameters), Equation (2) can be construed as an RT density function. Applying this assumption to the above example, this implies that categorizers will take longer to verify that a scene is a “dune” than to verify that it is “hilly.” This connection between average RT and average number of processing cycles gives us a first quantitative handle on the experimental literature in cognitive psychology. In the next section, we describe how SLIP can explain a large chunk of the so-called “basic-level” literature.

2. Basic-level literature

In Experiment 7 by Rosch et al. (1976), participants were taught the names of 18 objects at three levels of categorization: subordinate (e.g., “Levis,” “Macintosh”), basic (e.g., “pants,” “apple”), and superordinate (e.g., “clothes,” “fruit”) levels. These objects belonged to one of six possible nonanimal taxonomies: “musical instruments,” “fruits,” “tools,” “clothes,” “vehicles,” and “furniture.” In a verification task, subjects were shown a category name followed by a stimulus picture, and had to determine whether they matched. On average, categories at the basic level were the quickest to be verified, and categories at the subordinate level were the slowest to be verified [see also Murphy and Smith (1982), Hoffmann and Ziessler (1983), Jolicoeur, Gluck, and Kosslyn (1984), Murphy and Brownell (1985), Murphy (1991), Tanaka and Taylor (1991)]. This is the first of many verification experiments that demonstrated a superiority at a basic level of abstraction. The SLIP model is highly adapted to predicting basic-levelness, that is, the average speed of categorization at various levels of abstraction in a verification task.

Gosselin and Schyns (2001b) compared the predictive power of SLIP with that of four other basic-level measures of categorization: the context model [Medin and Schaffer (1978) modified by Estes (1994)], the category feature-possession model [Jones (1983)], the category utility model [Corter and Gluck (1992)] and the compression measure model [Pothos and Chater (1998, submitted)]. They used data from the empirical work of Corter, Rosch et al. (1976), Mervis and Crisafi (1982), Murphy and Smith (1982), Hoffmann and Ziessler (1983), Corter, Gluck, and Bower (1988), Lassaline (1990), Murphy (1991), Tanaka and Taylor (1991), Gosselin and Schyns (1998), Johnson and Mervis (1997), as well as from three novel experiments using computer-synthesized 3-D artificial objects. They found that SLIP “led the pack” by a large margin, predicting 88% of this (ordinal) data set, while predictions of this data set by the other models ranged from 64%, for the category utility model, to 35%, for the context model.

A critical aspect of basic-levelness is that it optimizes a number of indexes of performance. Convergence of all of these is crucial to establish a preferred categorization
level, even though verification speed is the most commonly used. It was thus important for Gosselin and Schyns to show that SLIP is not limited to model category verification, and they did demonstrate this; however, this goes beyond the scope of this chapter. The interested reader is referred to Gosselin and Schyns (2001b), where most of what we have described so far was originally presented. Next, we give a new twist to this relatively old story. We look at the way categorization occurs through time at the atomic level. This is an aspect of the question that has been completely neglected in the categorization literature.

3. Discrete processing cycles

The SLIP model proposes that we apprehend the world via discrete processing cycles, as discussed earlier. This is in clear disagreement with the apparent continuous nature of our everyday experience of time. However, our research team and others [e.g., VanRullen and Koch (2003b), VanRullen, Reddy and Koch (2005), Ward (2003)] have recently gathered compelling evidence that our visual world does in fact “tick.” Perhaps the most direct evidence comes from a recent study carried out in our laboratory using a powerful experimental technique called “Bubbles” [Gosselin, Lapalme, and McCabe (in preparation)]. We describe this experiment in some detail in the following section, which can also be seen as offering a “Bubbles primer”; this technique is at the heart of several other experiments reported later in this chapter.

3.1. A Bubbles primer

“Bubbles” is a generic procedure that can reveal the information that drives a measurable response [Gosselin and Schyns (2001a)]. Six decisions, or answers to six questions, are required in order to set up a Bubbles experiment: (1) what is the stimulus set? (2) in which space will the stimuli be generated? (3) what is the unit of sampling? (4) what is the observer’s task? (5) what are the observer’s possible responses? and (6) is the analysis per observer, or per group of observers? [Gosselin and Schyns (2005)]. Next, we discuss each of these decisions in the context of Gosselin et al. (in preparation).

3.1.1. Stimulus Set?

In a Bubbles experiment, the stimulus set is crucial, because it critically bounds what will be tested. Gosselin et al. used 640 natural scenes equally divided into eight categories (“city,” “highway,” “road,” “mountain,” “field,” “beach,” “house,” and “forest”). The scenes were gray-shaded images with a resolution of 128 × 128 pixels (subtending 2.8 × 2.8° of visual angle). The overall energy of the scenes was normalized. These scenes were presented for a duration of 150 ms. Generally speaking, the larger the stimuli set,
the better the *Bubbles* solution should be. A large stimulus set will tend to prevent observers from adopting strategies that are atypical of natural processing.

3.1.2. *Stimulus generation space?*

The choice of a proper stimulus generation space is one of the most important decisions when setting up a *Bubbles* experiment. Gosselin et al. searched the temporal dimension with a resolution of 120 Hz. We briefly describe other stimulus generation spaces in Section 5.1.1.

3.1.3. *The samples?*

At this stage, two important decisions have been made and the search can almost begin. In the search, information is sampled from the setup space, and the next decision to make concerns the unit of sampling. This unit depends on a number of factors, including the stimuli, the nature of the search space and the task to be performed. Here we varied the contrast of the scene through time. The contrast of the scenes presented by Gosselin et al. was modulated by a vector of Gaussian white noise low-passed with a Butterworth filter at 30 Hz (to prevent flicker fusion). Accuracy was maintained at 50% correct by adjusting the maximum contrast of the scenes using a gradient descent algorithm.

3.1.4. *The task?*

At this stage, the sampling procedure has been fully specified. Another important decision is that of the task. In Gosselin et al., it was a categorization task. Participants had to put the various scenes into the proper categories. Several tasks can and have been used with the same stimulus set.

3.1.5. *Response?*

The response is an interesting parameter of a *Bubbles* experiment, because the technique is, in principle, sensitive to any measurable dependent variable. “In Gosselin et al., observers pressed labeled keys corresponding to the eight scene categories and the accuracy of these responses constituted the dependent variable [see also, e.g., Bonnar, Gosselin and Schyns (2002), Gibson et al. (in press), Gosselin and Schyns (2001a), Schyns, Bonnar, and Gosselin (2002), Vinette and Gosselin (2002)]. The response latencies of such keypress responses [e.g., Schyns, Bonnar and Gosselin (2002), Smith, Gosselin and Schyns (2004)] and electroencephalographic (EEG) activity [e.g., Schyns et al. (2003), Smith, Gosselin, and Schyns (2004)] have been used as dependent variables in other experiments. Other dependent variables could include the firing rate of single cells, functional magnetic resonance imaging (fMRI) signal, galvanic skin responses, plethysmograph responses, and eye movements. To the extent that *Bubbles* is essentially an empirical tool, it is
useful to record as many different responses as possible. It is difficult to predict before the experiment how responses will correlate with the parameters of the search space.

3.1.6. Observers?

Depending on the objectives of the research, different types of observers can interact with the sampled stimuli. Gosselin et al. used human observers. Each performed about 3600 trials. Bubbles experiment have also been performed with human patients that had brain lesions and animals, as described later in Section 5.1.1.

3.2. Fossilized discrete processing cycles

Once the research experiment has been designed and the data have been collected, the analyses can be performed. The goal is to isolate a subspace of information that correlates with the measured response(s). Typically, a multiple linear regression on the samples (explanatory variable) and the response (predictive variable) provides this information. For the study by Gosselin et al. (in preparation), this analysis was reduced to summing up all the filtered noise vectors that led to a correct response in a first vector – the correct vector – and then adding all filtered noise vectors, irrespective of accuracy, in a different vector – the total vector. Then the correct vector was divided by the total vector, element by element. The result was a vector giving us the probability that if a temporal slot was sampled, a correct answer would be reached in this experimental setting, i.e., a time classification image. The left-hand column in Figure 2 illustrates the outcome of this analysis on one subject.

The most striking feature of this plot is the presence of a pulsation in the range of 7 Hz (i.e., within the \( \alpha \) bandwidth). We believe that this is a fossilized discrete processing cycle.

In a related vein, Vinette, Gosselin, and Schyns (2004) used a space–time version of Bubbles in order to examine how, during a face-identification task, visual information is extracted from stimuli in the first 280 ms after their onset. They obtained a clear pattern of results: the eye on the left side of the image became diagnostic between 47 and 94 ms after the onset of the stimulus; after 94 ms, both eyes were used effectively. More relevantly, Vinette et al. were the first to observe a sinusoid pulsation in the range of 7–14 Hz (i.e., within the \( \alpha \) bandwidth) in the effective use of temporal information.

In order for time-locked classification images to reveal anything at all, time slots need to possess special characteristics. This requirement was satisfied in the two experiments just described, because the processing cycles they revealed are (1) regular and (2) in phase with stimulus onset\(^1\). To a SLIP categorizer, however, temporal slots do not have any special meaning, and thus we cannot learn anything about a SLIP categorizer.

\(^1\) SLIP postulates processing “spikes” rather than the “sinusoid” oscillations observed. However, given some phase uncertainty (e.g., modeled by the convolution with a Gaussian function), the former can be made to mimic the latter.
with time classification images. In the next section, we show how the data from a Bubbles experiment can be analyzed differently and give us information about a SLIP categorizer.

3.3. What can temporal bubbles reveal about a SLIP categorizer?

So far, the Bubbles procedure has only been modeled for the Linear Amplifier Model (LAM) observer in the spatial domain [Murray and Gold (2004)]. Here, we sketch an answer to the question: what can temporal bubbles reveal about a SLIP categorizer?

We assume that a SLIP detector can fail at time $t$, either because of noise or because no bubble reveals information. In the experiment by Gosselin et al. (in preparation), a sample either reveals or does not reveal information at a particular time slot with equal probability. The behavior of a SLIP categorizer during such an experiment can be described by replacing the $\gamma_i$ in Equations (1) and (2) by $\gamma'_i = 0.5\gamma_i$. Apart from this “slowing down,” we assume that the categorization process of a SLIP categorizer is unaltered by the sampling occurring during a Bubbles experiment. For a thorough discussion of this issue, see Gosselin and Schyns (2004).

What really matters to a SLIP categorizer then is the number of processing cycles during which information was revealed. Suppose, for example, that on a particular trial, one bubble falls on the first processing cycle, and another bubble falls on the fourth processing cycle. Information is thus revealed for a total of two processing cycles. During other trials, information will be revealed for a total of 1, 2, 3, or more processing cycles. And, for each of these information slots, it is possible to compute the probability of a correct response. We will call the resulting vector of proportion correct an information classification image. As we mentioned earlier, the cumulative probability that a SLIP strategy comprising $n$ sets of redundant detectors is completed at processing cycle $t$ or earlier – i.e., the expected information classification image of any given SLIP categorizer – is given by Equation (1). The meaning of $t$ in Equation (1) must, however, be slightly modified: “$t$th processing cycle” must be replaced by “$t$th revealed processing cycle.”

In the Bubbles experiment of Gosselin et al. (in preparation), information classification images can be estimated as follows: the number of processing cycles available to an observer on trial $t$ is proportional to the sum of the elements of the filtered Gaussian noise vector employed to sample visual information on that particular trial, weighted element by element by that observer’s time classification image (see Section 3.2). The right-hand column in Figure 2 displays the information classification images extracted using this procedure on the ten subjects studied by Gosselin et al. (in preparation). In theory, once we have extracted the empirical information classification

Fig. 2. The left-hand picture gives the time-locked classification images of one subject using a temporal version of the Bubbles technique. Note the oscillation (in the $\alpha$ bandwidth; a total of about one cycle in the 150-ms temporal window of the experiment). The right-hand picture shows the result of a different analysis of the same data, i.e., information classification images.
image of a human categorizer, we can best fit it to Equation (1) (with the $\gamma$, not the $\gamma'$), and the parameters that minimize the error are estimates of the SLIP categorizer closest to that particular human being. Before we can apply this scheme practically, however, a tremendous quantity of work will need to be done. For example, we will have to study just how identifiable a SLIP categorizer is given an information classification image.

4. The need for flexibility and a paradox

So far, our story has been, to a large extent at least, a success story. We now turn to limitations of the SLIP model and similar models and look at ways to overcome these limitations. We hope that this will point toward new directions for research in categorization unfolding through time.

4.1. Limited processing capacity

There is a long and venerated tradition of research on the topic of information processing capacities in the field of human cognition [e.g., Broadbent (1958)]. We will not dwell much on this vast literature here. It will be enough for our purpose to cite a few representative examples. Most of the experiments in this field demonstrate one way or another that human information processing capacities are far less impressive than what we would naïvely expect. In a seminal article, Miller (1956) showed that our short-term memory has a capacity of seven, plus or minus two, “chunks” of information. Similarly, “object tracking” experiments performed by Pylyshyn and colleagues have shown that we can only track four or five moving targets simultaneously [e.g., Sears and Pylyshyn (2000)]. The most striking demonstrations ever perhaps come from so-called “change-blindness” experiments. When observers are asked to detect important changes in a natural or an artificial scene, they are typically shown to be ridiculously poor at this [e.g., Most et al. (in press), Rensink, O’Regan, and Clark (1997), Simons (2000b), see the special issue of Visual Cognition; Simons and Levin (1997)]. In the related paradigm of “inattentional blindness” [Mack and Rock (1998), Simons (2000b)], observers are asked to perform a task that, unbeknownst to them, is a distraction task (e.g., to count the number of times that members of a team in white T-shirts pass a basketball). In a small portion of the trials, something different happens (e.g., a human dressed in a gorilla suit walks to the center of the scene, turns toward the spectators, beats its chest with its hands a few times, and walks away). Usually, less than one-fourth of the participants notice these odd occurrences [Mack and Rock (1998), Simons (2000a)].

As we have written, en passant, the SLIP model implements this psychological reality by having only a subset of all available detectors activated simultaneously. So, if we are blind to a large change occurring in a natural scene – say, an engine disappearing from the wing of an airplane – it is because not enough of the relevant detectors are
active to complete the verification on time, and it suffices to activate more of these relevant detectors to see the change.

4.2. The need for flexibility

The limited processing capacity of the cognitive system implies that a selection of information must occur. There is now a wealth of evidence that this does, in fact, happen. We review some of the most compelling empirical evidences for this in Section 5.2. As a preview, we now look at an experiment performed with the four artificial scenes shown in Figure 1 [Gosselin and Schyns (submitted)]. In a learning phase, all participants learned to categorize the four scenes at a general and at a specific level. At a general level, the so-called LUMI participants, learned to separate the four scenes into “flat” and “hilly” on the basis of luminance cues; and the so-called CHRO participants, learned to separate the same scenes into “grassy” and “sandy” on the basis of chromatic cues. At a specific level of categorization, both LUMI and CHRO participants learned to categorize the stimuli as either “field” (the combination of is_flat and is_grassy), “desert” (is_flat and is_sandy), “mountain” (is_hilly and is_grassy), or “dune” (is_hilly and is_sandy). In a testing phase, participants were instructed to categorize the scenes at their most specific levels (never at their general levels). Note that the specific categorizations are strictly identical in the groups; the groups only differ in dimension structuring at high-level categorizations. The conjunctive nature of the stimuli can be used to determine indirect effects of diagnosticity. In the context of the SLIP categorizer, Gosselin and Schyns predicted that CHRO observers would weight the chrominance dimension more heavily than the luminance dimension, whereas LUMI observers weighted the luminance dimension more heavily than the chrominance dimension. This would happen if each group was tuned to chromatic and luminance information to maximize their categorization potential. After a successful test, only on the luminance (vs. chrominance) dimension, the LUMI (vs. CHRO) group could already categorize the scene at a general level, whereas the CHRO (vs. LUMI) group could not.

This does not help participants much in this experiment, but in real life, putting an object in a category allows people to infer unseen features [e.g., Rosch (1978), Anderson (1990)]. Consider the example of Pi Patel, the main character in Life of Pi [Martel (2003)], on his raft with an unknown thing. At least two categorization routes of similar processing time can lead him to the same conclusion: “The tiger ‘Richard Parker’ is standing just in front of me.” The first route would initially verify that the

---

2 Gosselin and Schyns (submitted) adapted SLIP to predict the error patterns of subjects in this situation. In a nutshell, they used Equation (1) and corrected it for guessing. The average best-fit curves for the various types of responses are reproduced in Figure 1b. Importantly, observer groups assigned orthogonal weights to the luminance and chrominance dimensions (with the CHRO (vs. LUMI) group biased to the chromatic (vs. luminance) dimension, with an odd ratio of 6:4), even though categorizations at the specific level (the task to resolve) was itself unbiased to one or the other dimension.
thing possessed distinctive tiger marks (property a), then verify the specific eye color of Richard Parker (property b). The second route would perform the same property tests, but in the opposite order (eye color before distinctive tiger marks). Both routes lead to the same final outcome in the same amount of time: “This is Richard Parker, a tiger.” However, in the first route, the initial testing of the tiger marks provides an intermediate “tiger” categorization before the specific “Richard Parker,” allowing Pi Patel to react faster. This intermediate categorization arises from the generic knowledge that tigers have tiger marks. In this struggle to categorize a thing, perspective matters. “Richard Parker” would probably prefer that Pi Patel uses the second route, so that he could enjoy an effortless dinner. In contrast, evolutionary arguments would probably select the first route: a “tiger” is after all a man-killer.

Very little is known about how categorizations are embedded in real life. Gosselin and Schyns’ experiment suggests, however, that beyond the probabilistic preference for categories with high basic-levelness induced by their feature structure, we are biased for a sequence of feature tests. There is actually some indication of this in the face recognition literature: Liu, Harris, and Kanwisher (2002), after having conducted a magnetoencephalography (MEG) study, proposed that face perception should be divided into two stages: a first stage where the stimulus is categorized like a face (occurring in the first 100 ms) and a second stage, completed after about 170 ms, where the face is identified at an individual level. Sugase et al. (1999) have obtained similar results performing unicellular recordings in macaque. These authors showed that information contained in a face is used in a first rapid stage, where global information allows the discrimination between macaque and human face, and a second slower stage, where finer information is used in order to identify face or categorize facial expression.

4.3. Back to the paradox

We have seen how the limited processing capacities of humans necessitate flexibility, and how this flexibility has, to some degree, now been incorporated into categorization models. However, the above SLIP story and ones like it, for all their appeal, inexorably lead to a paradox: how can a categorizer know which detectors to turn on, before knowing what is out there? And, reciprocally, how can a categorizer know what is out there before turning the relevant detectors on? We devote the next few pages to ways to resolve this paradoxical situation.

5. Categorization as an iterative process

Our resolution of the above paradox is twofold: We argue that a subset of a SLIP categorizer’s – or any other categorizer’s – detectors is always activated, and, that the remainder is used in a flexible manner, informed by previously activated detectors. Many theoretical proposals in the visual recognition literature are in agreement with
this answer. We will review these proposals, with categorization always on our minds. We will also describe a portion of the empirical work that supports these theories.

5.1. Compulsory feedforward processing sweeps

Ullman (1984) was among the first to propose that object recognition is informed by feedforward compulsory processing sweeps (or visual routines as he called this theoretical construct). Some detectors – using the terminology of the SLIP framework – would always be activated and would thus allow for surprise, for unexpected things to be discovered. Of course, these detectors cannot fully categorize the visual scene; otherwise, we would be back to square one, paradoxically-wise. This compulsory feedforward processing sweep, however, can attract our attention – guide the activation of our flexible detectors – toward suspicious-looking, partially processed objects.

Numerous studies performed by Thorpe and his research group as well as by others demonstrate the capacity of human subjects to categorize a visual scene very rapidly. Assuming that flexibility and feedback require time, this rapid processing supports the claim that a compulsory feedforward sweep can perform relatively complex processings [e.g., Thorpe, Fize, and Marlot (1996), Delorme, Richard, and Fabre-Thorpe (2000), Fabre-Thorpe et al. (2001), VanRullen and Thorpe (2001a)]. It has been demonstrated, for instance, that human participants can categorize a natural scene flashed for as little as 20 ms with high accuracy (94% correct) [Thorpe et al. (1996)]. In such ultrarapid categorizations, a differential electrophysiological component distinguishes target from nontarget scenes around 150 ms following stimulus onset [Thorpe et al. (1996)]; this brain activity is also correlated with the subject’s decision about the status of the stimulus [i.e., target vs. nontarget; VanRullen and Thorpe (2001b)]. It seems likely that the mechanisms involved in ultrarapid categorization are purely feedforward and encapsulated [Fabre-Thorpe et al. (2001), VanRullen, Delorme, and Thorpe (2001), VanRullen and Koch (2003a)]. The performance of subjects in such tasks does not increase with training even for period as extensive as 14 days [Fabre-Thorpe et al. (2001)]. Furthermore, a purely feedforward biologically inspired neural network was shown to be sufficient to duplicate ultrarapid categorization performance in humans [VanRullen et al. (1998), Delorme and Thorpe (2001)].

Likewise, it was shown that ultrarapid categorization of natural scenes [Li et al. (2002)] and gender discrimination of faces [Reddy, Wilken, and Koch (2004)] are possible in the near absence of attention. The performance of human subjects in both of these tasks was shown to be unimpaired by a dual-task requiring attention. Using a battery of experimental tools, LeDoux and colleagues [e.g., Armony and Ledoux (2000)] have shown that fearful faces are processed in a fast and feedforward manner by the amygdala and can subsequently guide behavior. One question that remains unanswered by all these experiments concerns the nature of the information processed during these compulsory feedforward sweeps.
5.1.1. *The nature of the information processed during compulsory feedforward sweeps*

Adolphs et al. (2005) used the *Bubbles* technique (see Section 3.1.1) to pinpoint the nature of the information processed by the amygdala during a fearful- vs. happy-face discrimination task. They compared the information used effectively to recognize emotion through facial expression by SM – a female patient with rare, complete, bilateral damage restricted to the amygdala – with the information used effectively by 10 matched controls. Each subject underwent about 3000 trials. In a trial, one of four faces was transformed in an image generation space comprising three dimensions (i.e., the standard X- and Y-axis of the image plane, and a Z-axis representing spatial frequencies), and was sparsely revealed by randomly located Gaussian holes. This sampling procedure is illustrated in Figure 3a. Adolphs et al. then performed multiple linear regressions on the location of the Gaussian holes and the response accuracy, to determine which areas of the image generation space were used most effectively. These regression coefficients were Z-scored, thresholded at 1.65 ($p < 0.05$), and used to construct effective faces. The results are illustrated in Figure 3b. They clearly show that the amygdala is involved in the processing of high-spatial frequency eye information.

In another study using the same *Bubbles* search space, Bacon et al. (2003), using the task devised by Dehaene et al. (2001), examined which type of information maximizes conscious and unconscious priming. Their results indicate that conscious priming depends mostly on local features in the high-spatial frequency range, whereas unconscious priming depends mostly on global features in the lower spatial frequency range (see Figure 3c). This, as discussed in the next section, is perfectly consistent with recent theoretical proposals by Bullier (2001a,b) and Bar (2003).

5.2. *Flexible iterative processing sweeps*

The previous section reviewed the evidence for the existence of compulsory feedforward processing sweeps or, transposed into the SLIP framework outlined above, for the compulsory activation of some detectors. Compulsory feedforward sweeps fit marvelously well within the “standard” feedforward anatomical hierarchy of the visual system [Felleman and Van Essen (1991)]. It is easy to understand why this has led mostly to bottom-up and constructivist models of information processing in the brain [Marr (1982), Biederman (1987)]. Recent experiments, however, have seriously challenged this viewpoint by showing the crucial importance of top-down processing (see also Section 4.2). These studies have thus set the stage for new explanatory models comprising either simple top–down components, or complicated iterative loops. We present physiological evidences for the existence of iterative loops as well as three models that mix early compulsory feedforward processing sweeps with late iterative processing sweeps.

According to Di Lollo, Enns, and Rensick (2000), Bullier (2001a,b), and Bar (2003), the visual scene would be partially analyzed by rapid and direct projections from the
early visual areas to the higher visual areas. This would yield a top–down working hypothesis informing the ongoing ascendant analysis. This descendant modulation would reduce the number of possible solutions for a particular retinal stimulation. Di Lollo’s model is not precisely constrained anatomically, unlike the other two models. In Bar’s model, information is first projected to the prefrontal cortex and then comes back to the inferotemporal (IT) cortex; in Bullier’s model, information is sent to V5 and then comes back to V1/V2. In both of these models, the first sweep contains low spatial

Fig. 3. (a) This illustrates how Adolphs et al. (2005) sampled faces at five independent bands of spatial frequencies (one octave each, with cutoffs at 22.38, 11.19, 5.59, 2.80, and 1.40 cycles/deg), to yield sparsely revealed images whose integration resulted in the stimuli that subjects saw. (b) Average difference between the effective use of information by patient SM and by the 10 healthy controls. (c) Results of Bacon et al. (2003). The top row contains the diagnostic regions for the conscious condition (i.e., high-spatial frequencies) and the bottom row contains the diagnostic information for the unconscious condition (i.e., low-spatial frequencies).
frequencies. We have already discussed empirical evidence for this in Section 5.1.1. Bar’s and Bullier’s models are illustrated in Figure 4.

5.2.1. Empirical evidence for flexible and iterative processing sweeps

Many physiological and electrophysiological studies lend support to the existence of such iterative processes, implicating feedback or reentrant information. We review the four types of argumentation that have been put forward: (1) the anatomical importance of brain projections (i.e., size of the pathways revealed by tracing and autoradiography techniques) and their functional importance (i.e., revealed by measures of the functional effect of inactivating one cortical area) are not always highly correlated [Vanduffel et al. (1997)], (2) despite what the topology of the visual system suggests, the visual cortex is temporally compact [Girard, Hupe, and Bullier (2001), Hupe et al. (2001)], (3) the temporal characteristics of neuronal response support the idea that cortical areas are implicated in different visual analyses at different moments in time [Lamme and Roelfsema (2000), Lee et al. (1998)], and (4) low-level cortical areas like V1 can produce sophisticated responses incompatible with their classical function of simple features detectors [Lee et al. (1998)]. The latter three points will be described in more detail in the next section.

5.2.1.1. The visual cortex is temporally compact. Two conditions must be met for the responses of neurons in low-level cortical areas to be modified through time: some neurons in higher-level areas must be activated rapidly, and these areas must provide rapid feedback to low-level cortical areas [Bullier (2001b)]. A recent meta-analysis of studies that measured latencies of the visual response of neurons in different cortical areas
revealed a temporal hierarchy that diverges considerably from the anatomical (classical) hierarchy [Lamme and Roelfsema (2000)]. Furthermore, this meta-analysis showed that neurons in the medial temporal (MT) and frontal eye field (FEF) areas are activated as rapidly as V1 neurons (MT: minimum = 39 ms, mean = 76 ms; FEF: min = 43 ms, mean = 91 ms; V1: min = 35 ms, mean = 72 ms) and more rapidly than neurons located in areas as low as V2 (min = 54 ms, mean = 84 ms) and V3 (min = 50 ms, mean = 77 ms) [Bullier (2001b), Lamme and Roelfsema (2000)]. This is perfectly consistent with the models by Bar (2003) and Bullier (2001a). Numerous factors could contribute to this lack of correspondence between topology and latencies of activation. First, neurons do not receive all their inputs via the shortest possible paths. Second, propagation speed of visual information differs according to neuronal pathways: a well-known distinction exists between magnocellular (fast), parvocellular (moderate) and koniocellular (slow) pathways. Lastly, it is possible to bypass the lateral geniculate nucleus (LGN) through, for example, the superior colliculus and the pulvinar, and to directly feed the extrastriate cortex with visual information [Lamme and Roelfsema (2000)]. Concerning the conduction speed of top–down pathways, Girard et al. [(2001), see also Panzeri et al. (2001)] observed fast feedback from V2 to V1 (roughly 3.5 m/s). This speed is more than sufficient to allow for a very rapid influence of high cortical areas on lower ones.

5.2.1.2. A given cortical area is implicated in different analyses at different moments. The response of cortical neurons is not constant. Instead, it seems that cortical neurons participate in different analyses at different moments [Lamme and Roelfsema (2000)]. Modulations in neuronal responses across time have already been observed in the LGN [DeAngelis, Ohzawa, and Freeman (1995)], V1 [Ringach, Hawken, and Shapley (1997, 2003)] and IT [Sugase et al. (1999)]. Ringach et al. (1997), for example, have shown that while the V1 neurons receiving a direct input from LGN (layers 4Cα and 4Cβ) have a constant preferred orientation through time, the preferred orientation of neurons in subsequent layers (2, 3, 4B, 5 and 6) drastically changes over time. It appears to be impossible to explain modulations such as the one just described by using an exclusively feedforward model [Ringach et al. (1997)]. Instead, Lamme and Roelfsema (2000) proposed that a compulsory feedforward sweep of activation lasting about 100 ms is followed by horizontal (i.e., from the same cortical area) and top–down influence lasting about 200 ms.

5.2.1.3. Low-order cortical areas are responsible of sophisticated responses. Cortical neurons, even those of V1, are not simple detectors responding selectively to one particular feature of the visual scene. Some neurophysiological data [e.g. Lee et al. (1998)] show that V1 is capable of sophisticated responses comparable with those of Ullman’s (1984) visual routines and Marr’s (1982) computations. As we have just seen, V1 processes different kinds of information over the 40–350 ms post-stimulation period. Although the initial V1 response (40–60 ms) seems to amount to local feature detection, numerous evidence shows that subsequent responses (80–200 ms) depend on
contextual information and involve higher-level processing [Kosslyn et al. (1995), Lee et al. (1998)]. According to Lee et al. (1998), the time-course of the V1 response argues for its gradual involvement in more and more sophisticated computations, and for its implication in tasks as complex as figure-ground segmentation and objects recognition. In sum, V1 would not be a simple module used in the processing of local features but would rather be a high-resolution buffer used for all sorts of visual processing feats.

5.2.2. Deactivation studies

The most direct evidences for a top–down influence on low-level visual processing perhaps comes from so-called deactivation studies. Hupe et al. (2001) have demonstrated a significant feedback effect of MT on V1, V2, and V3 less than 10 ms after deactivation [see also Girard et al. (2001)]. A series of recent transcranial magnetic stimulation (TMS) experiments with macaques and humans have shown that top-down processing is necessary for visual consciousness [Walsh and Cowey (1998), Pascual-Leone and Walsh (2001), Ro et al. (2003)]. For example, Pascual-Leone and Walsh (2001) impaired the conscious perception of moving phosphenes produced by stimulating area MT with TMS and by stimulating area V1/V2 with TMS 5–40 ms later. The most plausible interpretation of this result is that MT activation is not sufficient to perceive moving phosphenes, and that this perception requires V1/V2 activation in order to provide a spatial context to the stimulation. Other cortical areas responsible for conscious perception would also suffer from an interruption in top-down communication between MT and V1/V2 [Bullier (2001a), Pascual-Leone and Walsh (2001), Pollen (2003)].

6. General discussion

We could not come to terms with the thought of concluding this chapter without having even attempted to incorporate these relatively novel considerations about the necessity of having both compulsory and flexible feature detectors into a unique categorization model. Although we will not describe a fully articulated model here, we will “daydream” about such a model within the SLIP framework.

The question is, “What is the optimal way to use a subset of \( s \) flexible feature detectors to put an unknown object into one or many categories, given \( H_n \), a subset of our entire category hierarchy at processing cycle \( t \), and \( C \), a set of compulsory feature detectors?” Our working idea is to apply Bayes’ theorem sequentially [for another example of sequential use of Bayes’ theorem in categorization, see Anderson (1990, 1991)] to estimate \( P(f_{i,t+1}) \), the probability of encountering feature \( i \) in the environment at processing cycle \( t+1 \), given all the elements that we have already listed and \( d_{i,n} \), the fact that we have or have not detected this feature either with compulsory or flexible feature at processing cycle \( t \). Bayes theorem warrants that

\[
P(f_{i,t+1}) = k^{-1}P(f_{i,t})P(d_{i,t} \mid f_{i,t}),
\]

(3)
where \( k \) is equal to \( \sum P(f_i)P(d_i|f_i) \). At processing cycle \( t+1 \), we will activate \( sP(f_{i,t+1}) \) flexible detectors of feature \( i \).

To illustrate the computation of the two main components of Equation (3), consider once more Pi Patel on his raft facing an unknown thing. During the first processing cycle, Pi Patel detects distinctive tiger marks on the unknown thing with one of his compulsory feature detectors. This implies that only the features found in the “tiger” branches of Pi Patel’s complete category hierarchy should be looked for at time \( t+1 \); there is absolutely no need to search features found in the “inanimate” branches or in the “all animals except tiger” branches. Fortunately for us, Pi Patel knows only two tigers: “Richard Parker – defined by the additional Richard Parker’s eye color feature – and Walt Disney’s rendition of “Shere Khan,” the tiger from Rudyard Kipling’s classic novel The Jungle Book – defined by the additional Shere Khan’s eye color feature. Assuming that both these features are detectable by unique flexible-feature detectors, we have \( P(f_i) = 0.5 \), with \( i = \{ \text{Richard Parker’s eye color, Shere Khan’s eye color} \} \). All the other flexible detectors should be given a probability of being part of the unknown object equal to 0. In the Bayes’ theorem, this probability function is called the prior.

Is that all we can derive from the first processing cycle? No, we can also gain information about what is not out there, based on both the flexible- and the compulsory-feature detectors that were activated but have remained quiet. Suppose, for example, that some detectors of Shere Khan’s eye color were activated during this first processing cycle but did not fire. Either the unknown thing does not possess the Shere Khan’s eye color feature, or it does, but the detectors failed to detect it. We have already mentioned this possibility (failure of detection) in Section 1 of this chapter: in fact, \( P(d_i|f_i) = \phi \). Let us suppose, for the sake of the present illustration, that in the present case, this quantity is equal to 0.5. And because no Richard Parker’s eye color flexible feature detectors were activated during processing cycle \( t \), its associated \( P(d_i|f_i) \) is given a value of one. This probability function is known as the likelihood in Bayes’ theorem.

Combining the prior and the likelihood as shown in Equation (3), we obtain probabilities of \((0.5 * 1)/(0.5 * 1 + 0.5 * 0.5) = 2/3\) for the Richard Parker’s eye color and \((0.5 * 0.5) / (0.5 * 1 + 0.5 * 0.5) = 0.1/3\) for the Shere Khan’s eye color. This probability function is called the posterior in Bayes’ theorem. Finally, the posterior is multiplied by \( s \) to set the activation level of the flexible feature detectors at processing cycle \( t+1 \). If Pi Patel could simultaneously activate 90 flexible feature detectors due to limited processing capacities (see Section 4.1), he would activate \( 2/3 \times 90 = 60 \) and \( 1/3 \times 90 = 30 \), for Richard Parker’s eye color and for Shere Khan’s eye color, respectively.

An entirely satisfying account of effective categorization through time would address two more points: (1) how evolutionary pressures promote the cohabitation of mandatory and flexible detectors; and (2) how evolutionary pressures select the fixed detectors. Recent work by Geisler and Diehl (2002, 2003) that combines Bayesian models of perception with Bayesian models of evolution suggests a promising research avenue. We intend to fully develop the categorization model outlined above, and add this type of evolutionary spin to it, in the near future.
References


Gosselin, F., J. Lapalme and E. McCabe (in preparation), The Sinusoidal Pulsation of Attention During Categorization.


Chapter 35

ADAPTIVE CATEGORIZATION AND NEURAL NETWORKS*

ROBERT PROULX AND SÉBASTIEN HÉLIE

Université du Québec à Montréal

Contents

Abstract 794
1. The problem of divergence 798
2. The solution: dual Hebbian/anti-Hebbian learning 799
   2.1. Stabilization 800
   2.2. Oscillation 801
   2.3. Linearity 801
   2.4. Additional properties of the learning rule 801
3. The Eidos model 802
4. The letter classification task 803
   4.1. Methodology 803
   4.2. Results 804
5. The problem of convergence 806
6. The solution: Unlearning 808
7. The letter-classification task revisited 809
   7.1. Methodology 809
   7.2. Results 809
8. Current trends: Elimination of spurious attractors 812
9. Conclusion 813
References 814

* The work reported in this chapter represents a decade of research in the Laboratoire d’Études en Intelligence Naturelle et Artificielle. The authors are in debt to Jean Bégin and Sylvain Chartier for some of the innovations reported.
Abstract

Connectionist neural networks are useful geometric models of categorization. In particular, when *Hebbian learning* is used in a recurrent neural network architecture, it results in a recurrent associative memory (RAM), which convert the stimulus space into a feedback subspace sufficient to categorize new stimuli. However, the exclusive presence of positive feedback in Hebbian learning provokes the divergence of the eigenvalues of the weight matrix. In this chapter, we review some work that improved RAM capabilities. First, we show that the addition of the *dual Hebbian/anti-Hebbian* learning rule induces convergence of the eigenvalues of the weight matrix. This new learning procedure’s properties were assessed by using the Eidos model [Bégin and Proulx (1996)] to simulate learning in a letter-classification task. Second, the Eidos model was further improved by the addition of the *learning/unlearning* procedure, which guarantees that only the correct eigenvalues converge. Third, we review a newer model [Chartier (2004)] that reduces the number of spurious attractors by using the nonlinearity of the transmission function.
Studies about inductive categorization have generated an important body of research, since the early days of empirical psychology\(^1\). While early models of categorization described by some researchers were symbolic instantiations of popular theories [Reed (1972), Medin and Schaffer (1978)], other researchers presented evidence in favor of a metric psychological space [Shepard (1987), Tenenbaum and Griffiths (2002)] and suggested the use of geometric models of cognition, in which objects are represented by figures (vectors) in an \(n\)-dimensional space. Because the postulated psychological space is geometrical, the similarity between mental objects in these models can be computed, which was impossible in first-generation artificial intelligence models [Russell and Norvig (1995)]\(^2\).

One of the most popular families of cognitive geometric models is that of connectionist neural networks [Feldman and Ballard (1982)], which are models inspired from biology, where the basic units are “artificial neurons\(^3\).” Neurons are simple linear integrators with incoming and outgoing connections. If the activation impulse from the incoming connections is at or above a predefined threshold value, the neuron fires a signal that is sent along its outgoing connections (see Figure 1).

Connectionist neural networks are arrangements of several neurons that form a network that can be completely described by three elements: an architecture, a transmission function, and a learning rule. While many different architectural models have been used by cognitive scientists, we restrict this discussion to neural networks with a recurrent architecture (see Figure 2). A recurrent network usually includes a single layer in which all neurons are connected to each other [Anderson et al. (1977), Hopfield (1982)]. The input is defined by assigning values to particular neurons (in a process called clamping). After the initial clamping, the activation spreads to every other neuron to form the output, which is then fed back to the network to become the new input.

Because a single layer is used, the transmission function is usually quasi-linear and has a saturation value at a pre-determined value. Equation (1) describes the transmission in a typical recurrent network.

\[
x_{i+1} = L(Wx_i + x)
\]

\[
L(x) = \begin{bmatrix} L(x_1), L(x_2), \ldots, L(x_n) \end{bmatrix}^T;
\]

\[
L(x) = \begin{cases} \gamma & x \geq \gamma \\ x & -\gamma < x < \gamma \\ -\gamma & x \leq -\gamma \end{cases}
\]

where \(x\) is the stimulus vector at time \(t\), \(W\) the weight matrix, \(L(\cdot)\) the saturation function and \(\pm \gamma\) are the output boundaries. For simplicity, \(\gamma\) can be set to equal 1 [Anderson

---

\(^1\) See Murphy (2002) for an extensive review of categorization research.  
\(^2\) One way of assessing the similarity between symbols in classical AI models is to explicitly code the similarities in the symbols’ properties. This is not a viable solution, however, because there would be a combinatorial explosion of similarity measures that would need to be coded.  
\(^3\) Artificial neurons are simplifications of their biological counterparts. In this chapter, artificial neurons are referred to as neurons.
Fig. 1. Schema of an artificial neuron.

Fig. 2. Architecture of a single-layered recurrent neural network.
et al. (1977)] or the linear part of the transmission function can be omitted, thus resulting in a step function (Hopfield 1982)4.

A simple way to train recurrent networks is to reinforce connections between neurons with co-occurring activation [Kohonen (1972)]. The resulting learning rule is described by

\[
\Delta W_t = \eta x_{[p]} x^T_{[p]}
\]

\[
W_{t+1} = W_t + \Delta W_t
\]

where \(x_{[p]}\) the stimulus vector after \(p\) iterations, \(\eta\) a learning parameter, and \(\Delta W_t\) the weight change at time \(t\). This type of learning is called Hebbian learning, and when it is used in a recurrent architecture, it results in a recurrent auto-associative memory (RAM).

RAM neural networks have many advantages over other types of connectionist networks. For instance, they are an implementation of the simplest possible cognitive geometric model [Kosko (1988)]. The learning algorithm only requires local information, and the desired state does not need to be available to the network (unsupervised learning). Also, the quasi-linearity of the learning rule makes its dynamic easy to understand with currently available mathematical tools.

Moreover, learning involving RAM has a built-in capacity to generalize: The geometrical interpretation of RAM is that stimuli are trajectories in a hyperspace. Each iteration lengthens and shifts the angle of the stimulus toward learned attractors, which are created by the learning rule and are contained in the weight matrix (the eigenvectors). Therefore, when a new stimulus differs slightly from a previously learned stimulus, their trajectories also differ slightly and they are likely to stabilize in the same attractor. Thus, generalization is a free byproduct in RAM.

The most important consequences of the generalization capabilities of RAMs are robustness and tolerance to noise. The latter’s justification is direct: a noisy stimulus can be thought of as a new stimulus. Therefore, noisy stimuli have trajectories similar to their noiseless counterparts. On the other hand, robustness is a form of generalization that results from massively distributed representations. Because computation in connectionist neural networks is performed at the subsymbolic level [Smolensky (1988)], the absence or deterioration of some units is not dramatic. Whereas, in symbolic models, the loss of a unit represents a complete loss of representation, the lost of a neuron gracefully degrades the representation in RAM [Rumelhart (1989)].

Another useful property of RAM is its biological plausibility. In addition to the graceful degradation argument mentioned earlier [Rumelhart (1989)], auto-associative learning is a formalization of Hebb’s principle, which states that biological neurons that covariate share more synapses [Hebb (1949)]. Furthermore, the recurrent architecture corresponds well with Hebb’s work on cell assemblies [Kaplan, Sonntag and Chown

4 Hopfield’s (1982) network transmission rule also does not include the addition of the stimulus vector at each time step.
(1991)]. A cell assembly is a highly interconnected group of neurons that reverberate to sustain activity. According to Kaplan and his colleagues, RAMs can be used to represent cell assemblies.

Finally, the weight matrix resulting from Hebbian learning is equivalent to linear regression. The eigenvectors of the weight matrix (the attractors) form a new basis for the stimuli, which is optimal according to the least-squares criteria [Kohonen (1989)]. Therefore, categorization resulting from auto-associative memories trained with Hebbian learning have the same properties as their two-dimensional counterpart, which are already well documented [Larsen and Marx (2001)].

1. The problem of divergence

Even though the simple Hebbian RAM described earlier is an elegant model of human cognition, a serious problem prevents its use: the exclusive presence of positive feedback. After each learning trial, the value of the weight matrix is increased [\( \Delta W \) from Equation (2) is always positive]. Therefore, the eigenvalues of the weight matrix, which represent the radius of the attractors, grow with no boundaries. This serious problem, and the absence of quick solutions to explain this, resulted in the departure of research away from Hebbian learning models towards other learning algorithms; of these, competitive learning models [Grossberg (1976), Kohonen (1982)] and gradient descent methods [Rumelhart and McClelland (1986)] seemed particularly promising.

However, while competitive learning models [Grossberg (1976), Kohonen (1982)] and gradient descent methods [Rumelhart and McClelland (1986)] avoided the problem of divergence, they lacked many of the advantages emerging from Hebbian learning models (see the previous section). Moreover, these learning procedures brought their share of new problems. For example, gradient descent methods are unfit for unsupervised learning. The state generating the desired output must be available to the learning rule. Also, if a new item is added to the stimulus set after learning has begun, all learned stimuli must be relearned by the network, which is time-consuming and not psychologically plausible. As for competitive learning, the number of categories depends on the number of units and is thus predetermined by the network’s architecture.

The unsatisfactory nature of the proposed solutions, which avoided the divergence problem instead of solving it, suggested that the solution might reside in a reformulation of the problem. This is why we first concentrated on the mathematical analysis of the divergence problem.

Suppose, for the sake of simplicity, that the Hebbian RAM presented earlier is trained by repeatedly selecting stimuli at random from a set composed of \( k \) orthonormal items:

\[
\{ \mathbf{x}_i; i = 1 \ldots k, \mathbf{x}_i \in \mathbb{R}^n \text{ and } \mathbf{x}_i^\mathsf{T} \mathbf{x}_j = 1 \text{ if } i = j, 0 \text{ otherwise} \}.
\]
If none of the units reached saturation during learning, the saturation function can be dropped from Equation (1), which can be rearranged as

$$x_{t+1} = (W + I)x_t.$$  \(4\)

Because the stimuli are orthogonal, the state of the network after \(n\) iterations following the presentation of a stimulus vector \(x_i\) is

$$x_{[n]} = (W + I)^n x_i,$$

$$= (\lambda_i + 1)^n x_i,$$  \(5\)

where \(\lambda_i\) is the eigenvalue associated with \(x_i\). If the Hebbian learning rule (Equation 2) is applied to the output of Equation (5), Equation (2) becomes

$$\Delta W = \eta(\lambda_i + 1)^{2n} x_i x_i^T$$  \(6\)

and, because the stimuli are orthogonal, the addition of Equation (6) to the weight matrix \(W\) increases a single eigenvalue, \(\lambda_i\), by the value

$$\Delta \lambda_i = \eta(\lambda_i + 1)^{2n}. $$  \(7\)

Thus, the change in eigenvalue is a concave parabolic function with a vertex at \(-1\). Because the weight matrix is initially filled with zeros and the feedback is always positive, only the first quadrant must be considered in the analysis. Therefore, eigenvalues diverge and grow unbound, to infinity, at a rate following a power function (see Figure 3).

2. The solution: dual Hebbian/anti-Hebbian learning

In the preceding section, the effect of Hebbian learning on the eigenvalue spectrum of the weight matrix was shown to be a concave parabolic function centered at \(-1\). If the sign of the function is changed, the result is a convex parabolic function centered at \(-1\) (a reflection according to the abscissa; see Figure 3). This function qualifies as anti-Hebbian learning and decreases to minus infinity following a power function in the fourth quadrant. It is obvious from Figure 3 that Hebbian and anti-Hebbian learning curves cancel each other, but if the vertex of the anti-Hebbian function is slightly shifted to the right, the resulting function converges and learning eventually stops (remember that only the first quadrant must be considered). This is the idea behind dual Hebbian/anti-Hebbian learning, which is expressed mathematically by

$$\Delta W = \alpha x_{[p]} x_{[p]}^T + \beta x_{[n]} x_{[n]}^T, \text{ with }$$

$$\Delta \lambda = \alpha(\lambda_i + 1)^{2p} + \beta(\lambda_i + 1)^{2n},$$  \(8\)
where $\alpha > 0$ and $\beta < 0$ are, respectively, the Hebbian and anti-Hebbian learning parameters, and $p$ and $n$ are the number of Hebbian and anti-Hebbian iterations.

The properties of this learning procedure have been further explored by Proulx and Bégin (1990). In particular, three properties must be demonstrated in order to confirm the desirability of the proposed solution: it must stabilize, it must not oscillate, and it must remain linear. Stabilization is the property that confirms the convergence of the learning algorithm. As a result, the point of convergence and its necessary and sufficient conditions must be established.

The second property, oscillation, is related to the behavior of the eigenvalues as they approach the point of convergence: It must be shown under which conditions the eigenvalues reach the point of convergence and stop (instead of erratically moving around it).

Finally, linearity is a necessary condition for all the previous analysis. Therefore, the output boundaries of the transmission function ($\gamma$) must be set out of the reach of the iterating function in order to show that necessary and sufficient conditions have been established.

2.1. Stabilization

In order to stabilize (converge), the eigenvalue’s growth must have a root in the first quadrant ($\lambda > 0$) where the first derivative is negative. These conditions are satisfied when the
Hebbian portion of learning is greater than the anti-Hebbian portion, and the number of Hebbian iterations is smaller than the number of anti-Hebbian iterations. More precisely,

\[ \alpha > -\beta, \tag{9} \]

\[ p < n. \tag{10} \]

When these conditions are satisfied, it can be demonstrated that the convergence point \((\lambda_c)\) is

\[ \lambda_c = \left( -\frac{\alpha}{\beta} \right)^{1/(2(n-p))} - 1. \tag{11} \]

### 2.2. Oscillation

To prevent \(\lambda\) from oscillating around \(\lambda_c\), the eigenvalue must approach the point of convergence \((\lambda_c)\) slowly. Formally, this is represented by

\[ \left. \frac{\partial \Delta \lambda}{\partial \lambda} \right|_{\lambda_c} \geq -1. \tag{12} \]

A necessary and sufficient condition for Equation (12) to hold is that

\[ \alpha \leq \frac{\beta^{(2p-1)/(2n-1)}}{(n-p)^{(2(n-p))/2(n-1)}}. \tag{13} \]

### 2.3. Linearity

To maintain linearity, and all results presented in Equations (8)–(13), \(\gamma\) (the saturation point of the transmission function), must be chosen such that

\[ \gamma \geq \left( -\frac{\alpha}{\beta} \right)^{n/(2(n-p))}. \tag{14} \]

### 2.4. Additional properties of the learning rule

In addition to solving the problem of the eigenvalues’ growth, analysis of the dual Hebbian/anti-Hebbian learning rule revealed several other useful properties. First and foremost, the preceding analysis carried on the orthogonal case can be generalized to the correlated case without any other constraints on the parameters. Also, as in simple Hebbian learning, the feedback space resulting from learning is an optimal subspace of the stimulus space. In other words, the eigenvectors of the weight matrix form a new
basis for the stimulus set, which represents the stimulus space without significant loss of information\(^5\). Moreover, the learning rule imposes no constraint on the stimulus set: the same rule can be used without any outside task-related adjustment. Finally, unlike gradient-descent learning [Rumelhart and McClelland (1986)], the addition of new items to the stimulus set after the beginning of training redefines the feedback space accordingly, without need for relearning the entire stimulus set.

All these properties are adding psychological plausibility to the dual Hebbian/anti-Hebbian learning procedure. For instance, it is doubtful that humans categorize orthogonal and non-orthogonal stimuli differently. In fact, most empirical data suggest that the correlational structure is seldom learned (e.g. blocking) [Dennis and Kruschke (1998), Kruschke (2001)]. Also, the feedback space used to categorize new stimuli is a summary representation of encountered exemplars, which is in accordance with existing theories of categorization, for example, prototype theory [Reed (1972), Rosch and Mervis (1975)]. The striking correspondence between dual Hebbian/anti-Hebbian learning and empirical data, as well as the lack of need for outside adjusting or relearning of stimuli, make this learning procedure a more plausible model of human cognition.

3. The Eidos model

Now that an efficient learning rule for RAM has been developed, a simulation model is needed to intuitively illustrate this learning rule’s capabilities. The Eidos model [Bégin and Proulx (1996)] is a model of categorization in unsupervised neural networks, which uses the dual Hebbian/anti-Hebbian learning algorithm. The architecture of the Eidos model is similar to the usual single-layered recurrent network (Figure 2), except that connections inverting the signal have been added to illustrate the anti-Hebbian component of the learning rule (see Figure 4). The transmission function is similar to Equation (4), except that a short-term memory efficiency parameter has been added, representing a gradual degradation of the memory trace of the initial stimulus:

\[
x_{t+1} = L[(W + \psi I)x_t],
\]

where \(0 \leq \psi \leq 1\) is the short-term memory efficiency parameter and \(L[\bullet]\) the transmission function described in the second part of Equation (1).

The last element needed to completely define the Eidos model [Bégin and Proulx (1996)] is the learning rule, which is a variant of the dual Hebbian/anti-Hebbian rule described by Equation (8). The distinction is the addition of a long-term memory efficiency parameter affecting the weight matrix \(W\). The learning rule thus becomes

\[
W_{t+1} = \zeta W_t + \alpha x_{[p]} x_{[p]}^T + \beta x_{[n]} x_{[n]}^T,
\]

where \(0 < \zeta \leq 1\) is the long-term memory efficiency parameter, \(\alpha > -\beta\) and \(p < n\).

\(^5\) This general property holds in every unbiased environment.
The ability of the Eidos model to categorize various stimuli has been assessed in many classification tasks, such as letter classification [Bégin and Proulx (1996)], icon recognition, and face recognition. To illustrate the Eidos model’s potential in more detail, we next discuss how this model was used in a letter-classification task.

4. The letter classification task

4.1. Methodology

Nine stimuli were used in the letter classification task: \{A, E, H, I, N, O, R, S, T\}. Each stimulus represented a letter that was coded by a 35 unit binary vector (see Figure 5). This subset of the alphabet was chosen because it illustrates a wide range of between-stimuli correlations. The smallest absolute correlation is 0.03 (R-S and R-T pairs) and the highest is 0.77 (H-N and I-T pairs). At the beginning of the learning phase, the weight matrix (W) was filled with zeros. The free parameters were set according to the constraints presented earlier, so that: \(\alpha = 0.0005\), \(\beta = -0.00025\), \(\psi = 0.9\), \(\zeta = 0.9995\), \(p = 5\) and \(n = 10\). With those values, the point of convergence is \(\lambda_c = 0.0718\), and the saturation limit was set to its minimum value (\(\gamma = 2\)).
In the recall phase, Gaussian noise was added to the training stimuli in order to test the ability of the Eidos model to account for generalized learning [Bégin and Proulx (1996)]. The noise level varied from 0 to 100%. Examples of noisy stimuli are shown in the right-hand side of Figure 5. The network’s performance was evaluated by its ability to classify noisy inputs in the same categories as the corresponding noise-free stimuli.

4.2. Results

Figure 6 shows the eigenvalue spectrum’s development during learning. First, the dual Hebbian/anti-Hebbian learning rule succeeded in controlling the growth of the eigenvalue spectrum. Even after 10,000 trials, the highest eigenvalue was smaller than 0.35 (in contrast with the diverging growth of the Hebbian components). Also, the correct stimulus space was slowly emerging: nine eigenvalues were standing out from the remaining eigenvalues, confirming that nine categories were being learned. Finally, Figure 6, Figure 6 is Plate 35.6 in the Separate Color Plate section also shows that learning was incomplete after 10,000 trials. When learning was completed, the nine eigenvalues that were standing out were all equal to 0.0718.

In the testing phase, noisy stimuli were presented to the network in order to evaluate its ability to filter noise. The performance of the Eidos model was error-free when a noise proportion of 25 or 50% was added to the stimuli. Moreover, the error rate with stimuli covered by 100% noise was only 6.2%. Figure 7 shows an example of noise filtering on an E with 100% added noise.

Another interesting result is the analysis of the eigenvectors. Because the eigenvectors are the attractors in the Eidos model [Bégin and Proulx (1996)], the good performance of the Eidos model can be explained in terms of eigenvectors. Figure 8 shows three
examples of eigenvectors. As can be seen from the figure, these eigenvectors extracted the diagnostic features necessary to classify the stimuli. For instance, the left-hand eigenvector had high activation on its sides and inhibited the top and bottom middle parts. This pattern distinguished \{A, H, N, O\} from the remaining letters of the stimulus set. The middle eigenvector displayed high activation on all edges, with a preference for the top and bottom parts (which were excluded from the left-hand eigenvector). This eigenvector was especially important to partition the \{I, T\} and \{A, H\} subsets. Finally, the right-hand eigenvector displayed a preference for diagonals, which were contained in the \{N, R, S\} subset. Taken together, these rules are sufficient to classify any of the learned letters.

The letter-classification task illustrated the capacity of the Eidos model very well [Bégin and Proulx (1996)]. First, the network was able to learn a set of correlated patterns without domination from the first eigenvalue [Anderson et al. (1977)]. Second, the feedback space that emerged from the learning phase was sufficient to correctly classify stimuli even with up to 100% Gaussian noise.

Of particular interest was the analysis of the eigenvectors. As Figure 8 suggested, the eigenvectors developed in the letter-classification task acted as distributed categorization rules. Also, these eigenvectors represented a new basis for the stimulus set, which was an almost 4:1 compression. In cognitive psychology, it has been argued that one of
the many roles of categorization is cognitive economy [Reed (1999)]. There is no doubt that RAMs perform this economy.

5. The problem of convergence

The early success of the Eidos model with a wide range of stimuli (see Figure 9 for examples), highlighted its ability to classify highly correlated patterns (human faces), even with missing information. However, the convergence of the eigenvalue spectrum had an unexpected drawback: Figure 6 clearly showed that nine eigenvalues stood out from the remaining eigenvalues, but these eigenvalues eventually reached $\lambda_c$ and learning stopped (which is the definition of convergence). However, the remaining 26 eigenvalues, while relatively small, were positive. Therefore, if the number of learning trials was increased, these spurious eigenvalues would eventually reach $\lambda_c$, and the result

![Fig. 7. Example of noise filtering.](image)
Fig. 8. Three eigenvectors resulting from using the Eidos model in the letter-classification task.

Fig. 9. Examples of stimuli in a learning experiment using the Eidos model. Top panel: icons; bottom panel: human faces.
would be random behaviors of the network. Of course, the experimenter can decelerate the eigenvalues’ growth when the desired feedback space has been developed. This can be easily achieved by constraining the values of the memory efficiency parameters \( \zeta, \psi \) to be negative functions of the learning trial number. This solution is not satisfying, however, especially in the context of cognitive psychology, because it is difficult to justify that some category is eventually learned well enough, and that learning should, in the extreme case, cease.

6. The solution: Unlearning

A better way of controlling the growth of spurious eigenvalues is to use an unlearning procedure: After standard learning has occurred, negative feedback can be sent to the model to diminish spurious eigenvalues. The simplest way to accomplish this without adding complexity to the model is to reverse the parameters of the dual Hebbian/anti-Hebbian learning procedure. Figure 10 illustrates the resulting function and its dynamics.

The algorithm for the learning/unlearning procedure is as follow: First, a stimulus is randomly chosen from the stimulus set. Second, a noise vector is added to the chosen stimulus. Third, the dual Hebbian/anti-Hebbian learning rule is applied to the resulting noisy stimulus. Fourth, a second noise vector, with the same noise level as the first noise vector, is randomly generated. Fifth, the reverse dual Hebbian/anti-Hebbian learning rule is applied to the second noise vector. The four parameters of the reverse dual Hebbian/anti-Hebbian learning rule are not constrained to be the same as those

Fig. 10. The dual Hebbian/anti-Hebbian learning rule with reverse parameters. The arrows represent the order in which the learning/unlearning procedure must occur.
used in the standard dual Hebbian/anti-Hebbian learning model. The five preceding steps are repeated until the system stabilizes (typically after 10,000 trials).

7. The letter-classification task revisited

Some of the following simulations were first presented by Proulx et al. in 1996.

7.1. Methodology

The learning/unlearning procedure was applied to the letter-classification task (see Figure 5, right-hand side, for noise stimuli used). The noise level was set to 50% and the free parameters were $\alpha = \{0.005, 0.0025\}$ and $\beta = \{0.0025, 0.00125\}$, where the first value was used for learning and the second value for unlearning. The memory efficiency parameters’ effect $\{\psi, \zeta\}$ was neutralized by setting their value to unity. The number of iterations was the same for learning and unlearning, that is $p = 5, n = 10$.

7.2. Results

Figure 11 (Figure 11 is the Plate 35.11 in the Separate Color Plate section) shows the eigenvalues developed by the Eidos model [Bégin and Proulx (1996)] using this new learning procedure. As can be seen, the correct feedback space was developed and learning

![Fig. 11. Eigenvalue spectrum developed using the learning/unlearning procedure with the Eidos model. Eigenvalues are displayed in decreasing order.](image)
converged after 10,000 trials. Also, as learning continued after convergence of the correct
number of eigenvalues, spurious eigenvalues tended to decrease with practice, and were
negative after 10,000 trials. A negative eigenvalue acts as a vacuum that folds the hyper-
space by drawing its corresponding eigenvector toward the origin. As a result, stimuli pre-
sented to the neural network are converging faster toward eigenvectors associated with
meaningful eigenvalues. Hence, negative eigenvalues are no longer spurious; they are
helpful. Therefore, even if the number of learning trials grows to infinity, the spurious
eigenvalues do not interfere with the meaningful, correct ones. On the contrary, a longer
training period results in more pronounced negative eigenvalues, which is helpful.

Now that the desirability of the learning/unlearning procedure has been assessed, fur-
ther studies have been conducted to determine the ratio of unlearning to learning trials that
is necessary to obtain negative, useful eigenvalues. Figure 12, Figure 12 is Plate 35.12 in
the Separate Color Plate section shows the eigenvalue spectrum developed after 10,000
learning trials in the letter-classification task, as a function of the frequency of unlearn-
ing. First, the correct feedback space was developed, despite the value of the ratio of
unlearning to learning trials. Secondly, when the neural network was trained by using the
same number of learning and unlearning trials, the spurious eigenvalues were kept at zero.
A ratio higher than one resulted in negative spurious eigenvalues, which, as argued ear-
lier, are helpful. On the other hand, a ratio smaller than one produced positive spurious
eigenvalues, which are harmful to the network’s performance.

Fig. 12. Eigenvalue spectrum developed after 10,000 learning trials in the letter-classification
task. The eigenvalues are displayed in decreasing order and the frequency is expressed as the ratio
of the number of unlearning trials to the number of learning trials.
Now that simulations have shown that the correct feedback space is developed independently from the ratio of unlearning to learning trials, and that this ratio must be higher than one in order to generate helpful, negative eigenvalues, it remained to be determined whether the order of learning and unlearning trials affected the network’s performance. Figure 13 shows the eigenvalues developed after 10,000 learning trials in the letter-classification task. The unlearning to learning ratio was set to one, and an entry of 1000/1000 in Figure 13 (Figure 13 is Plate 35.13 in the Separate Color Plate section) means that a series of 1000 learning trials was followed by a series of 1000 unlearning trials. As can be seen, the sequence did not affect the development of the eigenvalue spectrum. Therefore, the experimenter can conveniently perform all the learning trials first, and then perform all the unlearning trials (instead of alternating between the two types of trials).

The preceding section presented simulations exploring the learning/unlearning procedure. As could be seen, this procedure was successful in providing a solution to the convergence problem, with only two restrictions: (1) there must be at least as many unlearning trials as learning trials and, (2) the learning trials must precede the unlearning trials. Assuming that these restrictions are met, the correct feedback space is developed, and the sequence of learning/unlearning trials does not affect the neural network’s performance.

Another interesting aspect of the unlearning procedure is that vectors used in this second phase basically consist of noise. From a psychological perspective, certain theories

![Fig. 13. Eigenvalue spectrum developed after 10,000 learning trials in the letter-classification task. The eigenvalues are displayed in decreasing order and the pattern of unlearning represent the type of sequences used to train the network. For example, 10/10 signifies a series of 10 learning trials followed by a series of 10 unlearning trials.](attachment:image.png)
of REM sleep consider dreams to be conscious interpretations of noise that is present in neural synapses [Crick and Mitchison (1983)]. Also, REM sleep is considered to play an important part in the consolidation of learning [Smith and Butler (1982)]. Therefore, unlearning can be interpreted as a model of REM sleep and the whole learning/unlearning procedure is well suited to model human learning.

8. Current trends: Elimination of spurious attractors

At this point, RAMs have gained many desirable properties. First, the dual Hebbian/anti-Hebbian procedure enabled learning to converge toward the optimal feedback subspace. Therefore, noise is optimally filtered, and new stimuli are classified in the most likely categories. Second, the learning/unlearning procedure results in faster convergence of stimuli toward meaningful attractors.

Solutions to the convergence and divergence problems made RAMs functional. We are now ready to tackle problems that used to be out of the reach of RAMs. We describe two main problems that have been identified. First, all previous RAMs have only been able to handle binary inputs. Most things in the world are described by real-valued vectors (e.g., the RGB color system). In order to be useful, RAMs should be able to directly handle such stimuli. Second, because the feedback space of a RAM is a subspace of the stimulus space, some stimuli – which are unknown to the network – exit from this feedback subspace and converge to attractors that are outside the feedback space but still in the stimulus space. This phenomenon strongly impairs recall performance in biased environments. This is why further research is now being performed to solve these newly identified shortcomings.

A model recently proposed by Chartier (2004) aimed at solving these two problems. More precisely, Chartier proposed a new converging RAM able to learn real-valued stimuli while considerably reducing the number of spurious attractors. This reduction in the number of spurious recalls was achieved by using the nonlinearity of the transmission function in conjunction with static temporal difference learning [Sutton (1988)]. The transmission function in the new model is

$$x_{t+1} = \begin{cases} 
1, & Wx_t \geq 1, \\
(\phi + 1)Wx_t - \phi(Wx_t)^3, & -1 < Wx_t < 1, \\
-1, & Wx_t \leq -1,
\end{cases}$$

(17)

where $\phi$ is a general transmission parameter and the learning rule is described by

$$W_{[t+1]} = W_{[t]} + \eta(x_{[0]}x_{[0]}^T - x_{[t]}x_{[t]}^T)$$

(18)

where $\eta$ is the learning parameter and $x_{[0]}$ the initial stimulus. Together, these equations allow the model to learn gray-scaled stimuli and converge toward an unequal eigenvalue.
spectrum with fewer free parameters [Chartier (2004)]. The unequal eigenvalue spectrum allows the model to use the environmental biases to optimize its recall performance [Hélie (2004)]. The result is a smaller number of random vectors stepping out of the feedback space. For example, Figure 14 shows trajectories of random stimuli in three-dimensional RAMs trained using two stimuli. The corners of the cube represent the attractors. In the linear model (left-hand panel), all eight corners are attracting random stimuli. However, because the network was trained using two stimuli, only four corners should be attractors (the two stimuli and their complements). The remaining four corners are spurious attractors. The right-hand panel of Figure 14 shows trajectories of random vectors in the new model. As can be seen, all random vectors are attracted to only four corners, representing the two learned stimuli and their complements. Therefore, in this case, the proportion of spurious states diminished from 100% in the linear model to 0% in the new model.

9. Conclusion

The previous work performed on simple Hebbian learning has highlighted important flaws with this learning procedure – in particular, the divergence of the eigenvalues of the weight matrix. A simple local solution was proposed in the form of the dual Hebbian/anti-Hebbian learning rule, which provided stable convergence points for eigenvalues representing the feedback space [Bégin and Proulx (1996)].

Convergence of the eigenvalue spectrum, however, had an important drawback: spurious eigenvalues, which were not useful to represent the stimulus space, eventually grew to converge at the same level as eigenvalues associated with the feedback space. Again, a solution was proposed – the learning/unlearning procedure – that resulted in negative spurious eigenvalues.
Current progress is now being made with converging RAMs that use the nonlinearity of the transmission function to classify real-valued stimuli. This new model [Chartier (2004)] can develop real-valued eigenvectors, associated with an unequally converging eigenvalue spectrum, as a means to diminish the number of spurious recalls. Spurious attractors are responsible for suboptimal recall performance in biased environments, and resolving this issue will certainly widen the applicability of RAM.

References

This page intentionally left blank
Chapter 36

A GROUNDED MIND IN A ROBOTIC BODY

STEVAN HARNAD

Université du Québec à Montréal

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved
The five chapters in this section – by Poirier et al., Proulx and Hélie, Cangelosi, Robert, and McCabe et al. – interdigitate quite well on the important themes of embodied cognition.

Poirier et al. first point out that minds (and brains) have bodies, and that this is not only unlikely to be incidental, but also most of the things that minds can do, they do with their bodies. Yes, “pure” thinking – i.e., cognition – seems in and of itself to be a disembodied mental activity, conducted autonomously inside our heads (if we agree to forget that the brain, too, is a bodily organ!), without any signs of sensorimotor interaction with the world of objects, organisms, states, events, and properties that most of our thoughts are about. But surely whatever pure thinking does go on in our heads occurs in the service of our present and future doings in the world, and is grounded in our past doings. And all of our doings are mediated by our bodies, specifically by our bodily organs of sensing and acting (along with whatever extrasomatic extensions we have managed to add to them through technology).

Our real-time cognitive performance – the activity that is measured and mapped in brain-imaging studies of thought – is less interesting and revealing than our cognitive capacity – the things we are capable of doing with our minds, rather than just the things we actually happen to be doing at any particular time. (This is the same thing as Chomsky’s celebrated distinction between “competence” and “performance.”) Cognition is hence a skill (or rather, a set of skills), or a body of essentially procedural know-how that we have (which is partly inherited and partly acquired from experience and learning). And despite the frequently made distinction between skill and knowledge – between knowing how and knowing that – it is becoming more and more evident that all of cognition is really the acquisition, possession, and exercise of a skill – a largely sensorimotor, body-based skill – whether it be starkly sensorimotor, as in athletic, artistic and practical skills, or more subtly so, as in our skill in spelling, speaking, writing, analyzing, fact-finding, arguing, proving theorems, or designing scientific theories. For these are all things we do with our bodies; they all have both an input side and an output side; and they all (to varying degrees) operate upon or are otherwise about the world external to our bodies – not only the concrete world of physical objects, but the abstract world of formal and fancied objects, such as prime numbers and unicorns.

As such, perhaps a simple way to summarize the embodied cognitive science that Poirier et al. describe is that it is a branch of reverse engineering committed to explaining the causalfunctional basis of our cognitive capacity by working to design a sensorimotor system (a robot) that is able to pass the Turing Test – i.e., able to do everything a human being can do, and do it so fully and so well that a human being could not tell that it was not done by another human being [Turing (1950), Harnad (2004)].

Turing’s original test had not called for a body. In fact, it had explicitly excluded the body from the test – not because the body’s capacities were not relevant, but rather, so that the candidate’s capacity could be judged without our being prejudiced by its physical appearance. We can probably drop that coyness now, as we have all already seen (and loved) enough cinematic robots that were far closer to human beings than anything that
has been built in a real robotics lab to date. We are, if anything, too ready to be credulous about robots today, when they are still so far away from having Turing power.

Turing’s original test was purely abstract and verbal. The testing was all to be done via e-mail; the test could have been passed by a computer just doing disembodied symbol-manipulation (computation). But it is now becoming apparent that not only is the disembodied e-mail version of the Turing Test not powerful enough to test whether the candidate can do everything a person can do (for surely we can do more than just read and write), but a computer just performing disembodied computation is not powerful enough even to read and write about everything a person can do, because its symbols are not grounded in the actual sensorimotor capacity to do any of those things.

Both Proulx and Hélie’s chapter and Cangelosi’s chapter are concerned with how to give a cognitive system that sensorimotor capacity. It is the capacity to detect (via its sensors), recognize and do the kinds of things we are able to do with the kinds of things there are in the world. In other words, it is the capacity to categorize. Think of that world of objects, people, events, actions, states, and properties that we are capable of naming, describing, and discussing in our e-mails. To be able to do that by e-mail, we have to have seen and done at least some of those things ourselves, with our bodies, in the world.

Proulx and Hélie provide a model for the sensory part of that skill. The shapes that objects project on our sensory surfaces can be processed by neural networks (nets) that do what is called “unsupervised learning.” By repeatedly being exposed to patterns, they adjust their connections to one another in such a way as to be able, after a while, to sort the input patterns into categories. Proulx and Hélie have improved on the learning capacity of existing unsupervised nets, so as to increase their skill in categorization. But unsupervised learning can still only find the categories that the patterns “wear on their sleeves,” so to speak. The next step is nets that learn by trial and error, receiving corrective feedback on whether they were right or wrong.

Cangelosi’s nets are of this kind. He shows how the nets can not only learn categories through direct sensorimotor trial-and-error experience, but also, once they have learned some categories in this way, the hybrid sensorimotor/symbolic system (of which the net is only a part) can then acquire further categories in a radically different way, a way that is unique to human beings: by “hearsay.” The nets can be told – i.e., given a string of symbols consisting of the names of categories they have already learned the hard way – what the properties of the new category are. The only condition is that the names of the properties must already have been grounded (the old way); their new recombination in the new category is expressed by recombining their category-names (symbols) into a description of the new category.

The world that such a hybrid robotic/symbolic system opens up – leading all the way from direct sensorimotor capacity to symbolic e-mail capacity – is the one that Robert focuses on in his chapter on the role of reasoning in cognition and categorization. Logical inference is the symbolic (computational) activity par excellence, the closest we ever get to “pure” thinking: formal symbol-manipulation, governed by rules (algorithms) that are based on the shapes (not the meanings) of the symbols. Robert asks
where the “creative” forms of inference come from, in this mechanical process. Hybrid systems offer several possibilities. One of them is embodiment. The sensorimotor “shape” of the world can influence the “shape” of ideas, especially through analogy. The other is through sensorimotor learning itself, which is not abstract formal induction, but concrete, trial-and-error induction, by corrective feedback from the bodily consequences of categorization – and miscategorization.

There does remain, of course, the question of where successful creative hypotheses come from (i.e., not those that result from the mechanical application of a rule). No one is going to explain the mind of Einstein before we first explain the mind of the average person with an IQ of 100. But even in average cognition – for example, ordinary category learning by children or adults – there is the “credit/blame” assignment problem: How does the system know, when it categorizes successfully or unsuccessfully, what property is responsible for its immediate success or failure? How does it arrive at the correct hypothesis eventually – or ever, in sufficiently difficult cases? McCabe et al. ask a similar question in their chapter (describing it as a “paradox”). “How can a categorizer know which detectors to turn on before knowing what is out there? How can a categorizer know what is out there before turning the relevant detectors on”?

Some (like Chomsky) have thought that the answer might be that there is an innate affinity between the structure of our minds and the structure of the world: that we find our nontrivial categories by “remembering” them (because they are cognitive universals that our brains possessed all along). Not only does this put a heavy burden on evolution, however (or, alternatively, on the Big Bang), but it may not even be necessary. No one really knows how improbable it is to find or construct the right feature-detector on the basis of any given trial-and-error sample of sensorimotor experience with corrective feedback, especially given sufficiently powerful learning mechanisms. Pasteur may be right that chance is the only remaining element needed to favor a mind that is sufficiently prepared in advance with both direct sensorimotor experience and symbolic instruction.

References


This page intentionally left blank
Chapter 37

CONCEPT LEARNING AND NONMONOTONIC REASONING

PETER GÄRDENFORS

Lund University

Contents

Abstract 824
1. The role of concepts 825
2. Three kinds of cognitive representations 825
3. Learning in symbolic systems 826
4. Learning in connectionist systems 827
5. Conceptual spaces as a representational framework 827
6. The origin of quality dimensions 829
7. Properties and concepts 831
8. Prototypes and conceptual spaces 832
9. Learning in conceptual spaces 834
10. The role of similarity in learning 836
11. Nonmonotonic aspects of concepts 838
   11.1. Change from general category to subordinate 839
   11.2. Context effects 840
12. Conclusion 841
References 842

1 This chapter is an expanded and revised version of Gardenfors (2001).

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved
823
Abstract

Humans learn new concepts extremely fast. One or two examples of a new concept are often sufficient for us to grasp its meaning. Traditional theories of concept formation, such as symbolic or connectionist representations, have problems explaining the quick learning exhibited by humans. In contrast to these representations, I advocate a third form of representing categories, which employs geometric structures. I argue that this form is appropriate for modeling concept learning. By using the geometric structures of what I call conceptual spaces, I define properties and concepts. A learning model that shows how properties and concepts can be learned in a simple but naturalistic way is then presented. This model accounts well for the role of similarity judgments in concept learning. Finally, as an application, the concept representations are used to give an analysis of nonmonotonic reasoning.
1. The role of concepts

The fundamental cognitive role of concepts is to serve as a bridge between perceptions and actions. In the simplest case, where there is a direct coupling between perception and action, a concept can be seen as a *decision procedure* where the perception is categorized and the chosen category then determines a choice of action. This is the role of concepts in lower animals (and the only role behaviorists would accept). In humans, and many of the higher animals, cognition is dominated by other, more indirect modes that are not triggered by perceptions alone. In these cognitive processes, concepts have a function in reasoning and in acting that is more independent of perception.

We are, in general, not born with our concepts; rather, they must be *learned*. In other words, the decision procedures that connect perceptions and actions must be created with the aid of the agent’s experience. In order to be useful, the procedures should not only be applicable to known cases, but should *generalize* to new situations as well.

The cost of generality is an increase in the frequency of error. Decision procedures may be more or less successful. When a particular perception is sorted under a concept, it may be a miscategorization, which in turn will lead the agent to choose the wrong action. If the agent realizes that it has made a mistake, it should adjust the application rules for the concept that led to the error. This is another aspect of concept learning. In this chapter, I will outline how *conceptual spaces* [Gärdenfors (2000)] can be used to model some of the fundamental aspects of how we learn concepts, how we adjust them, and how we reason with them.

2. Three kinds of cognitive representations

Humans are extremely efficient at learning new concepts. After being presented with only a couple of examples, we are able to abstract the general content of a new concept. A central problem for cognitive science is how this learning process and the underlying representations should be modeled. There are currently two dominant approaches to these problems. The symbolic approach starts from the assumption that cognitive systems can be described as Turing machines. In this view, cognition is seen as being essentially *computation* involving symbol manipulation. The second approach is associationism, where *associations* between different kinds of information elements carry the main burden of representation. Connectionism is a special case of associationism that models associations using artificial neuron networks. Both the symbolic and the associationistic approaches have their advantages and disadvantages. They are often presented as competing paradigms, but since they are used to analyze cognitive problems on different levels, they should rather be seen as complementary methodologies.

However, there are aspects of cognitive phenomena for which neither symbolic representation nor associationism seems to offer appropriate modeling tools. In particular, the mechanisms of concept learning cannot be given a satisfactory treatment in any of these representational forms. Concept learning is closely tied to the notion of *similarity*,...
which has also turned out to be problematic to model in the symbolic and associationistic approaches.

In this chapter, I will advocate a third form of representing information, which employs geometric structures rather than symbols or associations. Information is represented by points, vectors, and regions in dimensional spaces. On the basis of these structures, similarity relations can be modeled in a natural way in terms of distances in a space. I shall call my way of representing information the conceptual form because I believe that the essential aspects of concept learning are best described using this kind of representation. The framework of this chapter follows the theory presented in my recent book Conceptual Spaces: The Geometry of Thought [Gärdenfors (2000)].

Before presenting the geometric approach to concept learning, I shall briefly discuss how learning is modeled in symbolic and associationist theories and point out why I find these models lacking when it comes to the acquisition and development of concepts.

3. Learning in symbolic systems

The central tenet of the symbolic paradigm is that representing and processing information essentially consists of symbol manipulation according to explicit rules. The content of a symbolic expression is an agent’s belief or thought. The different beliefs in a person’s cognitive states are connected via their logical or inferential relations. Thus, the symbols are manipulated without considering their semantic content.

The central bearers of the symbolic representations based on first-order languages are the predicates of the language. These predicates are supposed to be given to the system. However, a successful system must be able to learn radically new properties from its interactions with the world, and not only form new combinations of the given predicates. This has turned out to be an enigma for symbolic representations. The problem presents an essential question for a theory of concept learning: where do new predicates come from? Furthermore, even after an agent has learned a concept, the meaning of the concept very often changes as a result of new experiences. In the symbolic mode of representation, there has been no satisfactory way of modeling the dynamics of concepts.

The problems concerning the formation and development of predicates become most pressing when one scrutinizes the attempts within the symbolic tradition to explain inductive inferences and nonmonotonic reasoning. The latter will be the topic of Section 11. The most ambitious project to analyze induction during the twentieth century was that of the logical positivists. Inductive inferences were important for them, since such inferences were necessary for their verificationist aims. However, it became apparent that the positivists’ methodology led to serious problems in relation to induction. The most famous ones are Hempel’s (1965) “paradox of confirmation” and Goodman’s (1955) “riddle of induction.”

What is needed is a nonlogical way of distinguishing the predicates that may be used in inductive inferences from those that may not. When it comes to inductive inferences, it is not sufficient that the references of predicates exist out there somewhere; we must
be able to represent them in our minds if they are to be used in planning and decision making. In other words, what is needed to understand induction, as performed by humans, is a conceptualistic analysis of predicates. Consequently, induction should be seen as closely related to concept learning.

4. Learning in connectionist systems

For Locke and Hume, thinking consisted basically of the forming of associations between “perceptions of the mind.” Since their time, this paradigm has been developed by British empiricists, American pragmatists, and psychologists. In the last few decades, associationism has been revived with the aid of a new model of cognition – connectionism. Connectionist systems, also called artificial neuron networks (ANNs), consist of large numbers of simple but highly interconnected units (“neurons”). The units process information in parallel (in contrast to most symbolic models, where the processing is serial). There is no central control unit for the network, but all neurons “act” as individual processors. Hence connectionist systems are examples of parallel distributed processes [Rumelhart and McClelland (1986)]. The behavior of the network as a whole is determined by the initial state of activation and the connections between the units. The inputs to the network also gradually change the “weights” of the connections between units according to some learning rule.

Connectionist systems have become popular among psychologists and cognitive scientists since they seem to be excellent tools for building models of associationist theories. Artificial neuron networks have been developed for many different kinds of tasks, including vision, language processing, concept formation, inference, and motor control. The applications include several that were traditionally thought to be typical symbol processing tasks. In favor of ANNs, the connectionists claim that these models do not suffer from the brittleness of the symbolic models and that they are much less sensitive to noise in the input [Rumelhart and McClelland (1986), Clark (1993)]. Furthermore, the learning mechanisms provide an answer to the question of how new concepts can be discovered by a cognitive system.

What, then, are the drawbacks of using ANNs of the type described here for representing information? The most pressing problem is that such systems need a very large training set to learn the relevant structures. The reason ANNs learn so slowly is that they have a large number of learning parameters, in the sense that all the weights between neurons are treated as independent variables that are slowly adjusted when the system is trained. As I will argue, learning mechanisms are much quicker on the conceptual level.

5. Conceptual spaces as a representational framework

As a framework for representing information structures that are relevant for concept learning, I have proposed the notion of a conceptual space [Gärdenfors (1990, 1992,
A conceptual space consists of a number of quality dimensions. Examples of such dimensions are color, pitch, temperature, weight, and the three ordinary spatial dimensions. These dimensions are closely connected to what is produced by our sensory receptors [Schiffman (1982)]. However, there are also quality dimensions that are of an abstract, nonsensory character. In Gärdnors (to appear), the analysis is extended to functional concepts.

The primary function of the quality dimensions is to represent various “qualities” of objects in different domains. Since the notion of a domain is central to my analysis, it should be given a more precise meaning. To do this, I will rely on the notions of separable and integral dimensions taken from cognitive psychology [Garner (1974), Maddox (1992), Melara (1992)]. Certain quality dimensions are integral in the sense that one cannot assign an object a value on one dimension without giving it a value on the other. For example, an object cannot be given a hue without also being given a brightness value. Similarly, the pitch of a sound always goes along with a particular loudness. Dimensions that are not integral are said to be separable, as for example the size and hue dimensions. Using this distinction, the notion of a domain can now be defined as a set of integral dimensions that are separable from all other dimensions.

The domains form the framework used to assign properties to objects and to specify relations between them. The dimensions are taken to be independent of symbolic representations in the sense that we can represent the qualities of objects without presuming an “internal” language in which these qualities are expressed.

The notion of a dimension should be understood literally. It is assumed that each of the quality dimensions is endowed with certain topological or geometric structures. As a first example, take the dimension of time. In science, time is a one-dimensional structure that is isomorphic to the line of real numbers. If now is seen as the zero point on the line, the future corresponds to the infinite positive real line and the past to the infinite negative line.

Conceptual spaces are suitable for representing different kinds of similarity relations: the closer two objects are located in a conceptual space, the more similar they are. If it is assumed that the dimensions have a metric, one can talk about distances in the conceptual space. Such distances represent degrees of similarity between the objects represented in the space.

It is important to introduce a distinction between a psychological and a scientific interpretation of quality dimensions. The psychological interpretation concerns how humans structure their perceptions. The scientific interpretation, on the other hand, deals with how different dimensions are presented within a scientific theory. The distinction is relevant when the dimensions are seen as cognitive entities, in which case their structure should not be determined by scientific theories that attempt to provide a “realistic” description of the world, but by psychophysical measurements which determine how our concepts are represented.

A psychologically interesting example of a domain involves color perception. In brief, our cognitive representation of color can be described by three dimensions. The first dimension is hue, which is represented by the familiar color circle. Thus, the topological
structure of this dimension differs from the quality dimensions representing time or weight, which are isomorphic to the real line.

The second psychological dimension of color is saturation, which ranges from gray (zero color intensity) to increasingly greater intensities. This dimension is isomorphic to an interval of the real line. The third dimension is brightness, which varies from white to black and is thus a linear dimension with end points. Together, these three dimensions, one with circular structure and two with linear, constitute the color domain, which is a subspace of our perceptual conceptual space. This domain is often illustrated by the so-called color spindle (see Figure 1). Brightness is shown on the vertical axis. Saturation is represented as the distance from the center of the spindle. Finally, hue is represented by the positions along the perimeter of the central circle.

6. The origin of quality dimensions

It is impossible to provide a complete list of the quality dimensions involved in the conceptual spaces of humans. Firstly, some of the quality dimensions seem to be innate or developed very early in life. They are to some extent hardwired in our nervous system. This probably also applies to our representations of ordinary space. Since domains of this kind are obviously extremely important for basic activities like getting around in the environment, finding food, and avoiding danger, there is an evolutionary justification for

![Fig. 1. The color spindle.](image-url)
the innateness assumption. Humans and other animals that did not have an adequate representation of the spatial structure of the external world were disadvantaged by natural selection.

Quine (1969, p.123) notes that something like innate quality dimensions is needed to make learning possible. However, once the learning process has started, new dimensions can be added. An example of this comes from studies of children’s cognitive development. Smith (1989, pp. 146–147) argues that working out a system of perceptual dimension, a system of kinds of similarities, may be one of the major intellectual achievements of early childhood. . . . The basic developmental notion is one of differentiation, from global syncretic classes of perceptual resemblance and magnitude to dimensionally specific kinds of sameness and magnitude.

Two-year-olds can represent whole objects, but they cannot reason about the dimensions of the object. Goldstone and Barsalou (1998, p. 252) note:

Evidence suggests that dimensions that are easily separated by adults, such as the brightness and size of a square, are treated as fused together for children. . . . For example, children have difficulty identifying whether two objects differ on their brightness or size even though they can easily see that they differ in some way. Both differentiation and dimensionalization occur throughout one’s lifetime.

Learning new concepts is, consequently, often connected with expanding one’s conceptual space with new quality dimensions. For example, consider the (phenomenal) dimension of volume. The experiments concerning “conservation” performed by Piaget and his followers indicate that small children have no separate representation of volume; they confuse the volume of a liquid with the height of the liquid in its container. It is only at the age of about five years that they learn to represent the two dimensions separately. Similarly, three- and four-year-olds confuse high with tall, big with bright, etc. [Carey (1985)].

Still other dimensions may be culturally dependent. Take “time,” for example; in some cultures, time is conceived to be circular – the world keeps returning to the same point in time and the same events occur over and over again. In other cultures, it is hardly meaningful at all to speak of time as a dimension. A sophisticated time dimension, with the full metric structure, is needed for advanced forms of planning and coordination with other individuals, but is not necessary for an organism’s most basic activities.

The examples given here indicate that many of the quality dimensions of human conceptual spaces are not directly generated from sensory inputs. This is even clearer when we use concepts based on the functions of artifacts or the social roles of people in a society. Even if we do not know much about the geometry of these dimensions, it is quite obvious that there is some such nontrivial structure. A preliminary analysis of functional concepts is given in Gärdenfors (to appear).

Culture, in the form of interactions between people, may in itself generate constraints on conceptual spaces. For example, Freyd (1983, pp. 193–194) puts forward the intriguing proposal that conceptual spaces may evolve as a representational form in a community just because people have to share knowledge.
Finally, some quality dimensions are introduced by science. Witness, for example, Newton’s distinction between weight and mass, which is of pivotal importance for the development of his celestial mechanics, but which has hardly any correspondence in human perception. To the extent that we have mental representations of the masses of objects in distinction to their weights, these are not given by our senses but have to be learned by adopting the conceptual space of Newtonian mechanics in our representations.

7. Properties and concepts

The theory of conceptual spaces will first be used to provide a definition of a property. I propose the following criterion [Gärdenfors (1990, 1992, 2000)], where the geometric characteristics of the quality dimensions are utilized to introduce a spatial structure for properties:

Criterion P: A natural property is a convex region in some domain.

A set is said to be convex if, for all points \( x \) and \( y \) in the set, all points between \( x \) and \( y \) are also in the set. Criterion P presumes that the notion of “betweenness” is meaningful for the relevant quality dimensions. This is, however, a rather weak assumption which demands very little of the underlying geometric structure.

Most properties expressed by simple words in natural languages seem to be natural properties in the sense specified here. For instance, I conjecture that all color terms in natural languages express natural properties with respect to the psychological representation of the three color dimensions. It is well known that different languages carve up the color circle in different ways [Berlin and Kay (1969)], but all such carvings seem to be done in terms of convex sets [Sivik and Taft (1994)].

Properties, as defined by criterion P, form a special case of concepts. I define this distinction by saying that a property is based on a single domain, while a concept may be based on several domains.

The distinction between properties and concepts has been obliterated in both the symbolic and the connectionist representations that have dominated the discussion in the cognitive sciences. In particular, both properties and concepts are represented by predicates in first-order languages. However, the predicates of a first-order language correspond to several different grammatical categories in a natural language, most importantly those of adjectives, nouns, and verbs. The main semantic difference between adjectives and nouns, on the one hand, is that adjectives like red, tall, and round normally refer to a single domain and thus represent properties, while nouns like dog, apple, and town normally contain information about several domains and thus represent concepts. Verbs, on the other hand, are characterized by their temporal structure, i.e., they essentially involve the time dimension. Even though it is impossible to generalize completely, most basic verbs represent dynamic properties of domains.

Concepts are not just bundles of properties. The proposed representation for a concept also includes an account of the correlations between the regions from different
domains that are associated with the concept. In the [APPLE] concept, there is a very strong (positive) correlation between sweetness in the taste domain and sugar content in the nutrition domain and a weaker correlation between the color red and a sweet taste.

These considerations motivate the following definition of concepts:

**Criterion C:** A concept is represented as a set of convex regions in a number of domains together with information about how the regions in different domains are correlated.

A third relevant notion is that of a category. Sloman, Love and Ahn (1998, p. 192) write: “Concepts and categories are, to a large extent, flip sides of the same coin. Roughly speaking, a concept is an idea that characterizes a set, or category, of objects.” I take a categorization to be a rule for classifying objects. The rule is generated from the representation of one concept or a category of concepts (such as the color concepts).

In my analysis of concepts, I have tried to bring in elements from other theories in psychology and linguistics. On the surface, the kind of representation proposed in Criterion C is similar to the frames with slots for different features that have been very popular within cognitive science as well as linguistics and computer science. My definition is richer since a representation based on conceptual spaces will allow me to talk about concepts being close to each other and about objects being more or less central representatives of a concept. My model can be seen as combining frames with prototype theory, although the geometry of the domains will allow predictions that cannot be made in either frame theory or prototype theory.

The notion of a concept defined here also has several similarities with the image schemas studied in cognitive linguistics by Lakoff (1987), Langacker (1987), and others. Even though their representations are often pictorial, they are, in general, not careful to specify the geometric structures of the domains that underlie the image schemas. Holmqvist (1993) has developed a computational-friendly representation that combines image schemas with some aspects of conceptual spaces.

8. Prototypes and conceptual spaces

Next, I shall argue that criterion P derives independent support from the prototype theory of categorization developed by Rosch and her collaborators [see, for example, Rosch (1975, 1978), Mervis and Rosch (1981), Lakoff (1987)]. When natural properties are defined as convex regions of a conceptual space, prototype effects are indeed to be expected. In a convex region, one can describe positions as being more or less central. In particular, if the space has a metric, one can calculate the center of gravity of a region.

It is possible to approach the argument from the opposite direction too and show that if prototype theory is adopted, then the representation of properties as convex regions

---

2 A slightly more complex definition, also involving the context where a concept is used, is proposed in Chapter 4 of Gärdenvors (2000).
is to be expected, at least in metric spaces. Assume that some quality dimensions of a conceptual space $S$ are given, for example the dimensions of color space, and that we want to partition it into a number of categories, for example color categories. If we start from a set of prototypes $p_1, \ldots, p_n$ of the categories, for example the focal colors, then these should be the central points in the categories they represent. If we make the additional assumption that $S$ is a metric space, the information about prototypes can be used to generate a categorization. In order to see this, assume that $S$ is equipped with the Euclidean metric so that, for every point $p$ in the space, one can measure the distance from $p$ to each of the $p_i$'s. If we now stipulate that $p$ belongs to the same category as the closest prototype $p_i$, it can be shown that this rule will generate a partitioning of the space – the so-called Voronoi tessellation. An illustration of the Voronoi tessellation is given in Figure 2.

A crucial property of the Voronoi partitioning of a conceptual space is that the Voronoi tessellation based on a Euclidean metric always results in a partitioning of the space into convex regions [see Okabe, Boots and Sagihiara (1992)].

Thus, assuming that a Euclidean metric is defined on the subspace that is subject to categorization, a set of prototypes will generate a unique partitioning of the subspace into convex regions by this method. The upshot is that there is an intimate link between prototype theory and criterion P. Furthermore, the metric is an obvious candidate for a measure of similarity between different objects. In this way, the Voronoi tessellation provides a constructive geometric answer to how a similarity measure and a set of prototypes determine a set of categories.

A Voronoi tessellation based on a set of prototypes is a simple way of classifying a continuous space of stimuli. The partitioning results in a discretization of the space. The prime cognitive effect is that the discretization speeds up learning, as we shall see in the following section. The reason for this is that remembering the finite prototypes, which is sufficient to compute the tessellation once the metric is given, puts considerably less burden on memory than remembering the categorization of each single point in the space.

![Voronoi tessellation of the plane into convex sets.](image)
In other words, a Voronoi tessellation is a cognitively economical way of representing information about concepts. Furthermore, having a space partitioned into a finite number of classes means that it is possible to give names to the classes. However, psychological metrics are imprecise and often context-dependent. Consequently, the borderlines will not be exactly determined.

9. Learning in conceptual spaces

In the previous section, I presented the prototype theory for concepts. Even if prototype theory fares much better than the Aristotelian theory of necessary and sufficient conditions in explaining how people use concepts, the theory does not explain how such prototype effects can arise as a result of learning to use concepts. The theory can neither account for how new concepts can be created from relevant exemplars, nor explain how the extensions of concepts are changed as new instances of the same category are learned.

I now turn to how these aspects of learning can be handled with the aid of conceptual spaces. Learning a concept often proceeds by generalizing from a limited number of exemplars of the concept [see, for example, Reed (1972), Nosofsky (1986, 1988), Langley (1996)]. Adopting the idea that concepts have prototypes, we can assume that a typical instance of the concept is extracted from these exemplars. If the exemplars are described as points in a conceptual space, a simple rule that can be employed for calculating the prototype from a class of exemplars is that the $i$th coordinate $p_i$ for the vector $p$ representing the prototype is defined to be the mean of the $i$th coordinate for all the exemplars [Langley (1996) p. 99]. The equation for the coordinates of a prototype $p$ is thus $p_i = \frac{1}{n} \sum x_{ik}$, where $x_{ik}$ denotes the locations of the $n$ exemplars that fall under the concept associated with $p$. The prototypes defined in this way can then be used to generate a Voronoi categorization.

Applying this rule means that a prototype is not assumed to be given a priori in any way, but is completely determined by the experience of the subject. Figure 3 shows how a set of nine exemplars (represented as differently filled circles) grouped into three categories with the aid of the equation above generate three prototypical points (represented as black crosses) in the space. These prototypes then determine a Voronoi tessellation of the space.

The mechanism illustrated here shows how the application of concepts can be generalized on the basis of only a few examples of each concept. The additional information that is required for the generalization is extracted from the geometric structure of the underlying conceptual space that is necessary for the calculation of prototypes and for the Voronoi tessellation. In this way, conceptual spaces add information to what is given by experience.

---

3 It is possible that an exemplar will lie outside the Voronoi region assigned to the corresponding prototype. Such an exemplar would be an anomaly for the classification. This kind of case will normally lead to the introduction of a new concept, i.e., a finer partitioning of the conceptual space.
Furthermore, the concepts generated by such a categorization mechanism are dynamic in the sense that, when the agent observes a new item $x$ in a category, the prototype for that category will, in general, change somewhat, since the mean of the class of examples will normally change. For each dimension $i$, the effect $\Delta p_i$ on dimension $i$ of the stored prototype $p$ can be calculated as $\Delta p_i = (x_i - p_i)/(n+1)$, where $x_i$ is the $i$th coordinate of the newly learned instance and $n+1$ is the total number of instances (including the new one) in the class represented by the prototype $p$ [see Langley (1996), pp. 99–100].

Figure 4 shows how the categorization in Figure 3 is changed after the agent learns about one new exemplar, marked by an arrow, which belongs to one of the categories. This addition shifts the prototype of that category, which is defined as the mean of the exemplars, and consequently the Voronoi tessellation is changed. The old tessellation is marked by hatched lines and the old prototype is marked by a gray cross.

It should be noted that the learning mechanism outlined here leads to very quick learning. Given only a couple of examples of a category, an approximate prototype can be calculated. The mechanism also puts very little load on memory. The only elements the concept learner needs to remember are the prototypes of the categories and an estimate of the number of examples these prototypes are based on. The remembered prototypes can then be used to generate a Voronoi tessellation that decides which category an arbitrary object should belong to. And if a new example of a category is observed, the number of previous observations of a category will determine how much the prototype should be adjusted. The larger the number of previous observations, the smaller the effect of a new observation will become.

The proposed mechanism also explains some aspects of the overgeneralization that often occurs in concept formation. For example, when a child learns the word dog (or a bow-wow version of it), this notion may well be applied not only to dogs, but to wolves, horses, calves, etc. The reason is presumably that the child has only mastered a few
prototypes in animal space and these prototypes will be used to generate a partitioning of the entire space. Consequently, the child will overgeneralize a concept in comparison to its standard use. After all, communication functions better if the child has some word for an animal than if he or she has no word at all. However, when the child learns new prototypes for other animal concepts, he or she will gradually adjust an early concept to its normal use since the partitioning of the animal space will become finer.

The learning mechanism presented in this section is not the only possible one that builds on dimensional spaces. Another strategy is to group objects according to the relative distances between them. This results in clusters of objects that can be useful when studying concept formation processes. Within the field of machine learning, there is a flourishing tradition concerning algorithms for identifying clusters.

10. The role of similarity in learning

One feature that clearly distinguishes the conceptual level from the symbolic is that similarity plays a central role on the conceptual level. Judgments of similarity are central to a large number of cognitive processes. Similarity relations between objects or properties, for example, that ‘green’ is closer to ‘blue’ than to ‘red,’ can be represented by distances in conceptual spaces. The learning mechanism presented in the previous section, which is based on distances in conceptual spaces, would be cumbersome to capture in a nonarbitrary way by symbolic representations and it could only be modeled in a roundabout way on the connectionist approach [see, for example, Schyns (1991)].

---

4 See, for example, Michalski and Stepp (1983), Murtagh (1993), and Langley (1996).
The representations formed in artificial neuron networks tend to be difficult to interpret. One reason for this is that dimensions are normally not represented explicitly in connectionist systems. It is true that ANNs learn about similarities but, in general, they do so very slowly and only after an exorbitant amount of training. The main reason for this is the high dimensionality of ANNs [see, for example, van Gelder (1995), p. 371]. The sluggishness of the learning is a result of each connection weight’s being adjusted independently of all the others. In addition to this, the adjustments are normally made in very small steps in order to avoid causing instabilities in the learning process. The assumption that the connection weights learn independently is not realistic from a neuroscientific point of view.

Since the vectorial representations of conceptual spaces have a much lower dimensionality and hence many fewer parameters that must be estimated, learning different kinds of patterns can be speeded up considerably by exploiting the conceptual level. This is an aspect of the learning economy of categorization that is often neglected. In brief, using conceptual spaces as the representational framework facilitates learning in artificial systems. The geometric structure of a dimension functions as a constraint that will make learning more efficient than it would be, for example, in an unstructured artificial neuron network. Of course, the geometric structure must have some correspondence in the external world – otherwise, what is learned may be useless or even dangerous.

For example, an explanation of how stimulus generalization works may become complicated, if the associationist approach is adopted. Stimulus generalization is the ability to behave in a new situation in a way that has been learned in other similar situations. The problem is how to learn which aspects of the learning situations should be generalized. This is an enigma for both symbolic and associationist representations. On the other hand, when conceptual spaces are used, the stimulus is represented as being categorized along a particular dimension or domain. The applicability of a generalization can then be seen as a function of the distance from a prototype stimulus, where the distances are determined with the aid of an underlying conceptual space.

One way of making learning in ANNs more efficient is to build in structural constraints when setting up the architecture of a network. In other words, one can reduce the dimensionality of the learning process by making neurons dependent variables. However, adding structural constraints often means that some form of information about the relevant domains or other dimension-generating structures is added to the network. Consequently, this strategy presumes a conceptual level in the very construction of the network. For example, one can use the technique of principal components in ANNs, which exploits redundancies in input data [see, for example, Schyns (1991), pp. 471–472, Schyns, Goldstone and Thibaut (1998), p. 15]. If the dimensions of the input data are correlated, principal component analysis finds the number (determined by the user) of orthogonal directions in the dimensions of the input data that has the highest variation. Thus, the first principal component is the spatial direction in the data set that has the highest variation and which thus is the maximally “explanatory” dimension.

However, Balkenius (1996) presents a connectionist model of stimulus generalization based on multiscale representations that avoids many of the problems that other associationist models have.
In this manner, the user can help himself or herself to some dimensional information that brings the representation of the ANN close to that of a conceptual space.

11. Nonmonotonic aspects of concepts

Concepts play an important role in the generation of inferences. As Holland et al. (1986, p.180) put it: “To know that an instance is a member of a natural category is to have an entry point into an elaborate default hierarchy that provides a wealth of expectations about the instance.” In this section, I will show that the model of concepts presented above will generate expectations that result in nonmonotonic reasoning.

The deductive reasoning of traditional logic is monotonic in the sense that, when a proposition \( A \) can be inferred from a set \( S \) of sentences, then \( A \) can be inferred from any set that contains \( S \). However, everyday reasoning, which in general is based on assumptions about what is “normally” the case, is often nonmonotonic. For example, if I learn that Gonzo is a bird, then, relying on the presumption that birds normally fly, I conclude that Gonzo can fly. But if I obtain the additional information that Gonzo is an emu, I no longer draw this conclusion.

Nonmonotonic reasoning is one of the hottest topics within AI. However, most research efforts have been concentrated on finding the appropriate logical rules that govern nonmonotonic reasoning. This means that a propositional (symbolic) representation of the relevant knowledge is already presumed. I believe that in order to understand nonmonotonic reasoning one must delve below the symbolic level of representation.

An important point that is often overlooked is that information about an object may be of two kinds: propositional and conceptual. When the new information is propositional, one learns new facts about the object, for example, that \( x \) is a penguin. When the new information is conceptual, one categorizes the object in a new way, for example, \( x \) is seen as a penguin instead of as just a bird. It is important to notice that describing information as propositional or as conceptual does not mean that these kinds of information conflict with one another. On the contrary, they should be seen as different perspectives on how information is described.

The theory of nonmonotonic inferences has focused on propositions; hence, it has been seen as a nonmonotonic logic. However, the great majority of the examples discussed in the literature derive from the nonmonotonicity of concepts. For example, the default rules studied by Reiter (1980) and his followers are conceived of as inference rules, although a more natural interpretation of “defaults” is to view them as relations between concepts. For instance, when something is categorized as a fruit, it will also, by default, be categorized as sweet, even though it is well known that the category contains many exceptions that are not sweet.

---

838

Peter Gärdenfors

See Gabbay, Hogger and Robinson (1993) for a survey of some of the research areas. In Balkenius and Gärdenfors (1991), nonmonotonic reasoning is connected to representations in artificial neuron networks; in Gärdenfors and Makinson (1994), it is connected to expectation structures.
It may be argued that there is no harm done in focusing on the propositional side of nonmonotonicity since information about categorization can be quite naturally transferred to propositional information: categorizing \( x \) as an emu, for example, can be expressed by the proposition ‘\( x \) is an emu.’ However, this transformation into the propositional form tends to suppress the internal structure of concepts. Once one formalizes categorizations of objects by predicates in a first-order language, there is a strong tendency to view the predicates as primitive atomic notions and to forget that there are rich relations between concepts, which disappear when represented in the standard logical formalism. Indeed, the fact that the concept [EMU] is a subcategory of [BIRD] is often represented by an explicit axiom in the form of a universal sentence ‘for all \( x \), if \( x \) is emu, then \( x \) is a bird.’ However, if the structure of concepts were built into the predicates of the language themselves, such an axiom would be totally redundant because the region representing emus would be a subregion of that representing birds. In fact, the inclusion relations between the regions and their domains will, in a sense, become analytic in a conceptual space.

To be sure, the literature provides one well-known theory of nonmonotonic inferences that focuses on conceptual relations, namely inheritance networks [see, for example, Touretsky (1986), Makinson and Schlechta (1991)]. However, in the theory of inheritance networks, concepts are represented by (nonstructured) points without any structure, and their relations by two kinds of links: “is-a” and “is-not-a.” Since these links say nothing about the structure of concepts, this form of representation is far too meager to handle the relations between concepts that are exploited in inferences, both monotonic and nonmonotonic. In contrast, I submit that a theory of conceptual structure is necessary in order to understand different kinds of nonmonotonic inferences involving concepts.

As a challenge to any theory of nonmonotonic inferences, I would like to focus on the following two aspects of the nonmonotonicity of concepts.

### 11.1. Change from general category to subordinate

This is the most well-known nonmonotonic aspect of concepts. When we shift from applying a “basic” category (a term borrowed from prototype theory) such as [BIRD] to an object \( x \), to applying a “subordinate” category such as [EMU], we often give up some of the properties associated with the prototype of the basic category: normally a bird is small, sings, flies, and builds nests in the trees, while an emu has none of these properties.

If concepts are described with the aid of convex regions of a conceptual space in accordance with Criterion C, such prototype effects are indeed to be expected. In a convex region, one can describe positions as being more or less central. For example, if color concepts are identified with convex subsets of the color space, the central points of these regions would be the most prototypical examples of the color.

Subordinate concepts may move away from the prototypes of the general concept and thus result in atypical properties. If the first thing I hear about the individual Gonzo is that it is a bird, I will, for lack of further information, naturally locate it in the conceptual space as a more or less prototypical bird, that is, at the center of the...
region representing birds as a subclass of animals. (The conceptual space needed to handle the concept of an animal may contain domains for shape, color, sound, behavioral variables, etc.) And in that area of the conceptual space, birds do fly. More precisely, almost all individuals that are located at the center of the bird region of animal space will also belong to the region of action space that represents animals that fly. However, if I then learn that Gonzo is an emu, I must revise my earlier concept location and put Gonzo in the emu region, which is a subregion of the bird region but presumably lies on the outskirts of that region. And in the emu region of action space, none of the individuals fly.

This simple example only hints at how the correlations between different regions representing properties in one domain and regions representing properties in another domain can be used to understand expectations and, consequently, nonmonotonic reasoning. For this analysis, the geometric structure of the regions in concepts is essential. A general inductive principle that underlies much human reasoning is that objects that are similar along some dimensions will be similar along other dimensions too. Such correlations will only be formulated in an ad hoc manner, if a symbolic representation of information is used where the similarity relations cannot be expressed in a natural way.

11.2. Context effects

Sometimes the mere context in which a concept is used may trigger different associations that lead to nonmonotonic inferences. Barsalou (1987, p. 106) gives the following example: “When animals is processed in the context of milking, cow and goat are more typical than horse and mule. But when animals is processed in the context of riding, horse and mule are more typical than cow and goat” [see also Roth and Shoben (1983)].

Another example of how the context affects the application of concepts is the following, from Labov (1973). He showed subjects pictures of various cup-like objects in order to determine how the variations in shape influence the names the subjects use. But he was also able to show that the suggested functions of the objects, for example serving as a container for coffee or for soup, influenced their naming. The result was that even if the “prototypes” of two concepts such as [CUP] and [BOWL] remain unchanged, the context may change the border between the concepts. The context makes certain properties more salient, which results in changes within the psychological space. Such changes may clearly have nonmonotonic effects on the use of the concepts.

The main effect of applying a concept in a particular context is that certain domains of the concept are brought into focus. In relation to the model based on Criterion C, this means that the context determines the salience of the domains. This results in a dynamic conceptual space, which in turn makes concepts and similarity judgments dynamic [see Smith and Heise (1992), pp. 242–248, Jones and Smith (1993), pp. 130–137].

Another effect of changes in context is that a change in the salience assigned to certain domains may result in a shift of the borders between different concepts. Once functionality was brought into focus in Labov’s study of the concept [CUP], the border between [CUP] and [BOWL] changed considerably. This kind of change can be
modeled mathematically with the aid of the Voronoi tessellations presented in Section 8. When the distance between two points in a space is determined, the relations between the scales of the dimensions is important. A change in the salience weight assigned to a dimension can be described as putting more weight on the distances between objects on that dimension, that is, magnifying the scale of the dimension.

For example, assume that we have a domain with two dimensions \( x \) and \( y \), and three prototypical points, \( p_1 \), \( p_2 \), and \( p_3 \), located as in Figure 5a. Assume that distances are given by the Euclidean metric. The Voronoi tessellation that is generated from these three points can then be determined. Now if the salience weight of the \( x \)-dimension is doubled, which means that the scale of \( x \) is multiplied by 2, the resulting Voronoi tessellation will change the categorization of some points in the space, as is shown in Figure 5b. For instance, the point \( q \) will belong to the category associated with \( p_1 \) in the first case, but to the category associated with \( p_3 \) in the second.

12. Conclusion

My aim in this chapter has been to model concept learning. I have pointed out some problems related to the symbolic and associationist accounts. Using conceptual spaces as an alternative mode of representation, I have presented simple learning mechanisms that can explain some of the most fundamental aspects of conceptual learning. These mechanisms are cognitively economical in that they demand very little of our memory capacities. I have also shown that the most common kinds of nonmonotonicity effects in reasoning can be modeled by exploiting the prototype structure and similarity relations of the concept representations.

In other words, I have shown that conceptual representations are also helpful when trying to understand how we communicate about concepts. In brief, conceptual representations can add significantly to our explanatory capacities when it comes to understanding cognitive processes, in particular those connected with learning, concept formation, (nonmonotonic) reasoning, and language understanding.

Fig. 5. Changes in Voronoi tessellation as a result of a change in the scale of the \( x \)-axis.
References


Berlin, B., and P. Kay (1969), Basic Color Terms: Their Universality and Evolution (University of California Press, Berkeley, CA).


This page intentionally left blank
Chapter 38

CATEGORIZATION IN SYMBOLIC DATA ANALYSIS

EDWIN DIDAY

Université Paris 9 Dauphine, Ceremade

Contents

Abstract 846
1. Introduction 847
2. Categories, concepts, and symbolic data 848
   2.1. From individuals to concepts 848
   2.2. Categories in a database 849
   2.3. From categories to concepts: reification of a category in a concept 849
   2.4. Sources of symbolic data 851
3. Symbolic data tables and their background knowledge, concepts, and categories 852
   3.1. Symbolic data tables 852
   3.2. Building a symbolic data table by reification of categories in concepts 852
   3.3. Description of concepts when the individuals are described by fuzzy data 853
   3.4. Adding conceptual variables, joining concepts, and the DB2SO module of SODAS 854
4. Modeling concepts by “symbolic objects,” with certain philosophical aspects 855
   4.1. Kinds of concepts and intuitive introduction of “symbolic objects” 855
   4.2. Modeling concepts with four spaces: “individuals,” “concepts,” “descriptions,” and “symbolic objects” 855
   4.3. Extent of concepts and symbolic objects 856
   4.4. Syntax of symbolic objects in the case of “assertions” 858
   4.5. Extent of a symbolic object 858
   4.6. Concepts: Four approaches 858
5. Tools for symbolic objects 859
   5.1. Order between symbolic objects 859
   5.2. Finding a unique description for a concept: “T-norm of descriptive generalization” 859
   5.3. Finding several descriptions for a concept 860
   5.4. Dissimilarities between concepts 861
   5.5. Finding prototypes from a concept 861
6. Underlying structures of symbolic objects 861
   6.1. A generalized conceptual lattice 861
   6.2. Mathematical framework of a symbolic data analysis 863
Abstract

We take a computer sciences-based approach to “categories” and “concepts,” as they emerge from databases. Our aim is to show how these notions from cognitive sciences can enhance “knowledge discovery” and “data mining” in computer sciences. Large databases are now in widespread use in industry and public administrations. They contain units described by variables that are often categorical. It is then easy to construct categories whose extents are units of the database. Whereas categories can be exhaustively described, the description of “concepts” (considered as higher-level units) can be improved to an infinite degree. The reification of categories in “concepts” yields “symbolic data tables,” where the units are the concepts and the variables’ values are “symbolic.” Symbolic data, represented by structured variables, lists, intervals, distributions, and the like, are better able to retain the complexity of concepts and their internal variation, than standard data, which they generalize. Lower-level (i.e., individuals) and higher-level units (i.e., concepts) are distinguished by this model, with symbolic descriptions constituting the former and “symbolic objects” the latter. Like the units themselves, the theory becomes more general. The aim of Symbolic Data Analysis is to generalize knowledge discovery, data mining and statistics to higher-level units described by symbolic data. The underlying structure of “symbolic objects” is defined by Galois Lattices. Finally, the software SODAS from EUROSTAT, which reifies categories in concepts and can discover new categories from these concepts (among other possibilities), is presented.
1. Introduction

Databases, sometime huge, can now be found everywhere in order to manage companies or to provide input to decision support systems. They are made up of data tables where the units are described by variables and can be linked together by other data tables. For example, a database might comprise three data tables. The first one describes customers (i.e., the units) by the variables home town, age, income, sex, etc.; the second one describes different kinds of cars according to size, cost, color, age, etc.; the third one associates each customer from the first data table with a car from the second data table.

This chapter is based on the following 10 steps:

1. A database is assumed to be given.
2. A set of categories based on the variable values in the database is chosen by an expert. For example, from the Cartesian product age x gender, if we retain three levels for age (young if less than 35, middle-aged between 35 and 70, old if more than 70), we get six categories.
3. We associate each category with its “extent” in the database, which is defined by the set of units that satisfy the category. For example, a male customer 20 years old satisfies the category (male, young).
4. By analogy with the real world, where an object cannot be defined by “what it contains or anything that we can assert on it” (Aristotle trans. 1994), we define a set called “real world” which contains two levels of units (also inspired by Aristotle). The first one is composed of units called “individuals,” the second is composed by units called “concepts.” The concepts, considered as higher-level units, have an extent composed of a set of individuals [Arnault and Nicole (1662/1965)].
5. A class of reified individuals can be associated with the class of units that defines the extent of a category. This class of individuals is considered to be the extent of the concept that reifies this category.
6. A set of “descriptions” is defined by a vector of variable values associated with each unit of the database and another set of descriptions (generally larger) is associated with its reified individuals. For example, a unit is described in a data table of the database by its gender and age but the associated individual is described not only by gender and age but also by his or her height and weight.
7. A generalization process is defined, in order to associate a description with a subset of individuals.
8. The generalization process is applied to a subset of individuals in the real world that is considered to belong to the extent of a concept. The description produced by this generalization process is considered to be a description of this concept. We note that the generalization process must take account of the internal variation in the set of individuals that it describes. Therefore, it leads to a new kind of variables defined based on the set of concepts, as the value of such variables for a concept can be a single numerical or categorical value, but also an interval, a set of ordinal or categorical values that may or may not be weighted, a distribution, etc. Since these values are not purely numerical or categorical, they have been called “symbolic data.”
and these new kinds of variables are called “symbolic variables.” For example, assume \{Paul, Jim\} constitutes the extent of the category (male, young) reified in the concept called [YOUNG MEN]; if Paul’s age is 35, and Jim’s age is 30, and the generalization process produces Age(young men) = [30, 35], then the value of the symbolic variable “Age” for the concept [YOUNG MAN] is an interval (we could also choose a distribution of the ages).

9. We define a “symbolic data table” where the units are the concepts and the variables are the “symbolic variables” that describe them.

10. Symbolic Data Analysis (SDA) extends Exploratory Data Analysis [Benzecri et al. (1973), Diday et al. (1979, 1984), Lebart, Morineau and Piron (1995), Saporta (1990), Tukey (1958)], and data mining (clustering, factorial analysis, discrimination, rules discovery, decision trees, Kohonen mappings, etc.) from standard data to symbolic data and produces categories of concepts which can be reified in new concepts and so on. More generally, SDA is able to discover new knowledge from the knowledge defined by concepts that reify specific categories extracted from a database.

Since the first papers announcing the main principles of SDA appeared [Diday (1987a,b, 1989)], much research has been published, including Bock and Diday (2000), Billard and Diday (2003) and various publications that came out under the auspices of the International Federation of Classification Societies publications [e.g., Billard and Diday (2000), Diday (2002), Limam, Diday and Winsberg (2004), Verde (2004)]. Recent papers can also be found in the Electronic Journal of SDA (JSDA) at www.jsda.unina2.it/newjsda/volumes/index.htm.

2. Categories, concepts, and symbolic data

2.1. From individuals to concepts

In his chapter of the Organon (Aristotle trans. 1994) devoted to categories, Aristotle clearly distinguishes “first-order individuals” (such as “a horse” or “a man”), considered as units associated with individuals in the world, from “second-order individuals” (such as “the horse” or “the man” in general), taken as units associated with classes of individuals.

Starting from this idea, we define a first set called “set of real-world units” composed of two subsets: the set of units called “individuals” and the set of units called “concepts” (which correspond to the first- and second-order individuals, respectively, introduced by Aristotle). Following Aristotle, these units (he would say “substances”) are defined by their own existence and cannot be defined by “what they contain” or what we “can assert.” In other words, whereas what we can assert about a category can be strictly defined, any strict definition of an individual or of a concept destroys it. Hence, we can only approximate their definition by means of what we can assert. These assertions are based on a second set called “set of descriptions,” made up of “variables” able to approximately describe the units in the real world. These descriptions are
obtained by linking the units with variables of different kinds expressing quantities
(Jim’s height), relations (Jim is larger than Tom), qualities (Jim is black), places (Jim
lives in London), passions or experiences (Jim is burned), times, situations, actions,
behaviors, etc. Such a list of variables can never be exhaustive. In order to see that an
individual called “Toby” is a “dog” (i.e., matches the concept of [DOG]), we compare
the description of the concept [DOG] and the description of Toby with the same set of
variables. If the values of the variables match, Toby is considered to be a dog accord-
ing to this set of variables. The (never attained) complete description of a concept is
called its “intent,” while the set of individuals which match the intent is called the
“extent.” Using “universal ideas” for “concepts,” Arnault and Nicole (1662/1965)
brilliantly defined these notions in the framework of the Port-Royal School:

Now, in these universal ideas there are two things which are important to keep quite distinct: intent and
extent. I call the intent of an idea the attributes which it contains and which cannot be taken away from
it without destroying it. Thus the intent of the idea of a triangle includes, to a superficial extent, figure,
three lines, three angles, the equality of these three angles to two right angles, etc. I call the extent of
an idea the subjects to which it applies, which are also called the inferiors of a universal term, that being
called superior to them. Thus the idea of triangle in general extends to all different kinds of triangles.

2.2. Categories in a database

A first meaning of a “category” is given in natural language, where a “category” is a class
of similar units. A second meaning of a “category” is defined in philosophy by the quality
which can be attributed to an unit. Hence, Aristotle defined 10 categories (substance, quan-
tity, relation, quality, place, passion, time, situation, action, and behavior), and Kant defined
four classes of categories (quantity, quality, relation, and modality). In this chapter, we con-
sider Aristotle’s and Kant’s categories as kinds of “variables” which exhaustively define a
class of similar units called “individuals,” and considered as “first-order” individuals.

Any relation or logical expression that links units of a database with the values of a
given set of variables of the database constitutes a kind of description called a “cate-
gory” in this chapter. More specifically, a category is defined by a name and a set of
true or false properties linking the individuals of the real world and the variables. This
set constitutes a description called the category’s “intent” and the units of the database
which satisfy these properties are called its “extent.” The description of a category,
which defines it, is always exhaustive. For example, “Tall and heavy men” is the name
of a category defined by the description “Height greater than 2 meters” and “Weight
greater than 100 kg,” using the variables “height” and “weight.”

2.3. From categories to concepts: reification of a category in a concept

We can associate a concept with any category by assuming that the extent of the cate-
gory in the database is reified in an extent of this concept in the set denoted Ω of indi-
viduals of the real-world set (see Figure 1).
Therefore, the intent of the category (given by using the variables which define it), insofar as it is a description of its extent, constitutes a partial description of the concept that it reifies. This arises from the fact that this description can be completed by any description (using other variables) of the class of individuals that defines the concept’s extent. For example, the category named “young people,” whose description is “people less than 18 years old,” constitutes the partial intent of the concept [YOUNG PEOPLE] as the full intent of this concept may also contain many other descriptions concerning, for instance, the height and the weight of the people less than 18 years old in \( \Omega \). Going back to Arnault and Nicole’s example, the intent of the category whose name is “geometrical triangle” defined by “three lines, three vertices,” is simply an approximation of the definition of the concept of same name [GEOMETRICAL TRIANGLE], whose intent may also contain “figure, three angles, the equality of these three angles to two right angles,” etc.

Any description of a concept using variables of the database defines a category. If this description contains the necessary conditions satisfied by the concept, the reification of the extent of the category in the set of individuals \( \Omega \) contains the extent of the concept. For example, the concept [TALL MAN] can be described by “height more than 2 meters,” which defines a category by a relation concerning the variable “height.” The union of all the intents of such categories constitutes the approximate intent of this concept. Hence, the intent of a concept contains the intent of any such category, and its extent is contained in the extents of these categories (more specifically, in the intersection of their extents). In practice, the intent and extent of a concept in the real world is never exhaustively defined, as the number of variables can always grow to infinity. A definition can be always improved. Reverting back to Arnault and Nicole’s example, the category whose name is “figures with three angles” and whose description is “figure, three angles,” is simply an approximation of the concept [GEOMETRICAL TRIANGLE] whose intent may also contain “three lines, the equality of these three angles to two right angles,” etc. We can see that the intent of the concept [GEOMETRICAL TRIANGLE] contains the intent of the category “figures with three angles” and its extent is contained in the set of figures with three angles which is the extent of the category.
2.4. Sources of symbolic data

There are many sources of symbolic data. First, one can find the extent of a concept
given by the extent of any category, any categorical variable, or any Cartesian product
of categorical variables contained in a given database. Hence, the intent of a category
can first be defined as a value or a Cartesian product of categorical variables, for exam-
ple, “female” or “female, more than 70 years old.” More generally, a category can be
defined by a structured query language (SQL) query (i.e., a first-order logical predicate)
of a given database. After that, it is easy to associate each category with its extent,
which is the set of units of the database that satisfies its intent, as defined by the SQL
query. It is possible to reify the category in a concept whose partial description is
defined by a description of the extent of this category, using other variables from the
database. For instance, in a “soccer database,” the variable “team” associates each
player with his or her team. Hence, from the category “French team,” we can obtain the
set of players in the database who play on the French team. We can then reify this cat-
egory in a concept called [FRENCH TEAM]. This concept can then be described by a
description of the set of soccer players on this team, with other variables as age, height,
salary, etc., contained in the database.

Another way to determine the extent of a new concept is to use a clustering process
(K-means, Kohonen maps, etc.) on a set of units described by a set of variables contained
in the database: each cluster can then be considered as the extent of a particular concept.

In both cases, symbolic data are then needed, in order to obtain a description of the
concept from the set of individuals that constitute its extent. For example, coming back
to the example of a soccer database, if we consider all the categories (i.e., all the teams)
defined by the variable “team” and we reify them in concepts, we obtain a data table
where the units are concepts (here, teams) described by new variables that are defined
on the set of concepts, instead of the initial variable in the database, which was defined
on the set of individuals (here, the set of players). The values of these variables must nec-
essarily take account of variation in the individuals within the extent of each category
(i.e., each team). For example, by examining the variations in the age, height, salary, etc.,
among the set of players on the French team, we obtain intervals or sets of values or dis-
tributions. For example, if François is a player on the French soccer team, we might have
\( \text{age(François)} = 20 \) for the initial variable “age.” But if we consider the concept “French
team players” as a unit, we need to use a new kind of variable defined on the set of con-
cepts (i.e., teams). If they are denoted by capital letters, we obtain \( \text{AGE(French team), Nationality) = \{[21, 32], (0.9 \text{ French}, 0.1 \text{ African})\} \), which means that the age of play-
ners on the French team ranges from 21 to 32 and that 90% of the players of this team are
French. We could also choose the interquartile interval or the distribution of the French
players’ ages as a value of the variable age. These new kinds of variables are called
“symbolic variables” and their values are called “symbolic data,” since they are not num-
bers on which the standard numerical operators can be applied. Hence, when a category
is reified in a concept, the concepts can be considered to be new individuals associated
with each row of a “symbolic data table,” where the columns are associated with each
symbolic variable. Each cell of this table contains the symbolic value taken by the concept associated with its row for the variable associated with its column.

Symbolic data may also be “native” in the sense that they result from expert knowledge (traffic accident scenarios, types of emigration, species of insects, etc.). They can also come from the probability distribution, the percentiles, or the range of any random variable associated with each cell of a data table, from time series (in representing each time series by the histogram of its values or in describing intervals of time), from confidential data (in order to conceal the initial data by reducing its accuracy), etc.

3. Symbolic data tables and their background knowledge, concepts, and categories

3.1. Symbolic data tables

Symbolic data tables constitute the main input of a SDA. They are defined as follows: columns in the input data table are “symbolic variables,” which are used in order to describe a set of units called “individuals.” Rows are called “symbolic descriptions” of these individuals because they are not merely vectors of single quantitative or categorical values. Each cell of this symbolic data table contains different types of data:

(a) Single quantitative value: for instance, if “height” is a variable and \( w \) is an individual: \( \text{height}(w) = 3.5 \).

(b) Single categorical value: for instance, \( \text{Town}(w) = \text{London} \).

(c) Multivalued: for instance, in the quantitative case, \( \text{height}(w) = \{3.5, 2.1, 5\} \) means that the height of \( w \) can be either 3.5 or 2.1 or 5. Note that (a) and (b) are cases of (c).

(d) Interval: for instance \( \text{height}(w) = [3, 5] \), which means that the height of \( w \) varies in the interval \([3, 5]\).

(e) Multivalued with weights: for instance, a histogram or a membership function. (Note that (a) and (b) are cases of (e) when their weights are equal to 1 or 0.) Variables can be:

(f) Taxonomic: for instance, “the color is considered to be ‘light’ if it is ‘yellow,’ ‘white,’ or ‘pink.’”

(g) Hierarchically dependent: for instance, we can describe the kind of computer that belongs to a company only if it has a computer; hence the variable “does the company have computers?” and the variable “kind of computer” are hierarchically linked.

(h) With logical dependencies, for instance: “if \( \text{age}(w) \) is less than 2 months then \( \text{height}(w) \) is less than 10.”

Many examples of such symbolic data are given in Chapter 3 of Bock and Diday (2000).

3.2. Building a symbolic data table by reification of categories in concepts

To recapitulate, one starts with a set of categories, which are reified in concepts whose description can be improved ad infinitum. Table 1 gives an example of data obtained
from a specific database where each unit is a housing unit for sale in a given district, which is considered as a category. Each unit is described by its kind of accommodation, its area, the area of its garden, its price, the number of rooms, the area of its living room, and the number of bathrooms. Each category is then reified in a concept representing a district. The units are reified in individuals whose partial description results from the description of these units. We use a generalization operator (defined in Section 5.2) applied to the extent of each district in order to get Table 2. Hence, in this table, each concept is a district described in a row by symbolic variables whose values express the internal variation in the initial variables among the housing units for sale in the district listed in the database. For example, in Paris 14 the prices range from 409 thousand euros to 599 thousand euros. So the “value” taken by the symbolic variable “Price” for the concept called [PARIS 14] is the interval [409, 599]. For the same concept, the symbolic variable “accommodation” is multivalued with weights; its value for [PARIS 14] is expressed in terms of the frequency of house selling (0.5) and flat selling (0.5). Table 3 provides other examples of symbolic variables in a symbolic data table describing four concepts.

3.3. Description of concepts when the individuals are described by fuzzy data

It can happen that individuals are described by fuzzy variables. For example, the variable “high income” is a fuzzy variable whose values range between 0 and 1, expressing for each individual how high his or her income is. For example, “high income”(Paul) = 0.9 means that the individual Paul’s income is high. In Table 3, each concept is described by the symbolic fuzzy variable “HIGH INCOME,” which expresses the internal variation in the fuzzy values of the variable “high income” taken by the individuals that make up its extent. Hence, in Table 3, “HIGH INCOME” (Concept 1) = {0.35} means that all the individuals of the extent of Concept 1 have taken the fuzzy value 0.35 for the variable “high income.” “HIGH INCOME”(Concept 2) = {0.32, 0.88} means that the fuzzy values taken by the individuals of the extent of Concept 2 for the variable “high income” range from 0.32 to 0.88. These values are 0.31, 0.46, or 0.72 for Concept 3. For concept 4, they vary between 0.2 and 0.3 with a frequency of 0.4, and take the value 1 with a frequency of 0.6.

Table 1

Each individual is a housing unit for sale; each category is a district

<table>
<thead>
<tr>
<th>Units</th>
<th>Category</th>
<th>Accommodation</th>
<th>Area</th>
<th>Garden</th>
<th>Price</th>
<th>Rooms</th>
<th>Living</th>
<th>Bath</th>
</tr>
</thead>
<tbody>
<tr>
<td>122</td>
<td>Paris 20</td>
<td>Flat</td>
<td>47.00</td>
<td>0.00</td>
<td>189</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>123</td>
<td>Paris 20</td>
<td>House</td>
<td>84.00</td>
<td>32.00</td>
<td>365</td>
<td>3</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>74</td>
<td>Paris 14</td>
<td>Flat</td>
<td>80.00</td>
<td>0.00</td>
<td>409</td>
<td>2</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>Paris 14</td>
<td>Flat</td>
<td>90.00</td>
<td>0.00</td>
<td>542</td>
<td>3</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>76</td>
<td>Paris 14</td>
<td>Flat</td>
<td>111.00</td>
<td>0.00</td>
<td>599</td>
<td>3</td>
<td>34</td>
<td>2</td>
</tr>
</tbody>
</table>
3.4. Adding conceptual variables, joining concepts, and the DB2SO module of SODAS

Once the symbolic data table which describes the concepts obtained from a generalization of the classes that are the extent of its categories has been built, other variables related to the concepts themselves, as opposed to the individuals, can be added. For example, in Table 2, we could add the number of inhabitants and the area of each district.

For example, the categorical variable “city” (whose values are London, Cambridge, etc.) defined on a set of inhabitants listed by a census, associates each inhabitant with the city where he or she lives. Then, with the categorical value “London,” we can associate the concept [PEOPLE LIVING IN LONDON] whose extent is the set of inhabitants of London. In another census, the variable “city,” defined on a set of schools listed by another census, associates each school with the city where it is located. Then, to the categorical value “London,” we can associate the concept [SCHOOLS IN LONDON], whose extent is the set of schools in London. It is then possible to model the concept [LONDON SCHOOLS AND PEOPLE] by uniting the symbolic descriptions of the categories [PEOPLE LIVING IN LONDON] and [SCHOOLS IN LONDON] without having to merge both census databases at the level of people or schools. In section 5, we will see how to induce these symbolic descriptions and symbolic objects from the database by a generalization process.

The SODAS software’s DB2SO module allows the possibility of (i) building a symbolic data table describing concepts, by means of a generalization process applied to the extents of categories obtained from a query to a database; (ii) adding specific variables to the concepts induced by the categories; and (iii) joining symbolic data tables containing different variables, in order to join concepts. For more details on DB2SO, see Stephan, Hébrail and Lechevallier (2000).

### Table 2
Each category is reified in a concept described by symbolic variables

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Accommodation</th>
<th>Surface</th>
<th>Garden</th>
<th>Price</th>
<th>Rooms</th>
<th>Living</th>
<th>Bath</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris 20</td>
<td>House (0.5),</td>
<td>[47, 84]</td>
<td>[0,32]</td>
<td>[189,365]</td>
<td>{1,3}</td>
<td>[31,32]</td>
<td>{1}</td>
</tr>
<tr>
<td></td>
<td>Flat (0.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paris 14</td>
<td>Flat (1)</td>
<td>[80,111]</td>
<td>[0, 0]</td>
<td>[409,599]</td>
<td>{2,3}</td>
<td>[24,34]</td>
<td>{1.2}</td>
</tr>
<tr>
<td>Vincennes</td>
<td>House (1)</td>
<td>[93, 93]</td>
<td>[0, 0]</td>
<td>[495,495]</td>
<td>{2}</td>
<td>[36,36]</td>
<td>{2}</td>
</tr>
</tbody>
</table>

### Table 3
A “symbolic data table”: each cell contains an example of “symbolic data” where the variable HIGH INCOME expresses the variation in fuzzy values

<table>
<thead>
<tr>
<th>Concept</th>
<th>HIGH INCOME</th>
<th>CITY</th>
<th>SOCIOECONOMIC GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept 1</td>
<td>{0.35}</td>
<td>{London}</td>
<td>{Service employees}</td>
</tr>
<tr>
<td>Concept 2</td>
<td>[0.32, 0.88]</td>
<td>{Paris, London}</td>
<td></td>
</tr>
<tr>
<td>Concept 3</td>
<td>{0.31, 0.46, 0.72}</td>
<td>{0.1 Manager, 0.6 Manual, etc.}</td>
<td></td>
</tr>
<tr>
<td>Concept 4</td>
<td>[(0.4) [0.2,0.3],[0.6] [1]]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Modeling concepts by “symbolic objects,” with certain philosophical aspects

4.1. Kinds of concepts and intuitive introduction of “symbolic objects”

We consider two kinds of “concepts”:
(i) “Concepts in the world,” such as a town, a region, a road accident scenario, a kind of unemployment. This kind of concept is defined by an “intent” and an “extent” which exist, have existed, or will exist in the real world.
(ii) “Concepts in our mind” [like the so-called “mental objects” explained by Changeux (1983)], which model in our mind concepts from our imagination or concepts from the world by their properties and a “way of finding their extent” (by using our sensors), and not the extent itself since there is no room for all the possible extents in our mind.

A “symbolic object” models a concept in the same way as our mind does it, by using a description $d$ (representing its properties), a matching relation $R$ (representing our sensors) between the description of a new individual and $d$, and a mapping $a$ using $d$ and $R$, in order to say whether the new individual belongs to the extent of the concept. For example, the description (in my mind), of what I call a car, its matching (following several sensors: eyes, ears, hands, etc.) with an object that I can see, hear, touch, etc., and a way of recognizing that a given object is a car by making a synthesis of these matches. Symbolic objects are defined more specifically in Section 5. As it is quite impossible to obtain all the characteristic properties of a concept and its complete extent, a symbolic object is simply an approximation of a concept and the problem of the quality, robustness, and reliability of this approximation arise. This important question will be discussed in Section 4.3.

4.2. Modeling concepts with four spaces: “individuals,” “concepts,” “descriptions,” and “symbolic objects”

In Figure 2, the “set of individuals” and the “set of concepts” constitute the “real-world” set, while the “modeled world” is made up of the “set of descriptions” which models individuals (or classes of individuals) and the “set of symbolic objects” which models concepts. We start with a “concept” $C$ whose extent denoted $\text{Ext}(C/\Omega)$ is known in a sample $\Omega$ of individuals. For example, if the concept $C$ is “insurance companies,,” and 30 insurance companies are among a sample $\Omega$ of 1000 companies, $\text{Ext}(C/\Omega)$ comprises these 30 insurance companies. Each individual $w$ of the extent of $C$ in $\Omega$ is described by using the mapping $y$ such that $y(w)$ describes the individual $w$. The concept $C$ is modeled in the set of symbolic objects by the following steps, as described in Figure 2.
(i) We generalize the set of descriptions of the individuals in $\text{Ext}(C/\Omega)$ with an operator $T$ in order to produce the description $d_C$ (which can be a set of Cartesian products of intervals and/or distributions). In Section 5, we discuss this operator in more detail.
(ii) The matching relation $R$ can be chosen in relation with the $T$ choice. For instance, if $T = \cup$ then $R_C = \subseteq$, and if $T = \cap$, then $R = \supseteq$ (see examples given in 6.1).
The membership function is then defined by $a_C(w) = \int y(w) \, R \, dC$, which measures the fit or matching between the description $y(w)$ of the unit $w$ and the description $d_C$ of the extent of the concept $C$ in the database. We can then define the symbolic object modeling the concept $C$ by the triple $s = \langle a_C, R, d_C \rangle$. In the next section, we will revisit this definition and illustrate it with several examples.

Knowing the extent $E_1$ of a concept in $\Omega'$ and the extent $E_2$ in $\Omega'$ of a symbolic object that models this concept, the aim of the learning process is to improve the choices of $T$, $R$, and $a$, until $E_1$ and $E_2$ are as close as possible. In Section 5.3, we expand on this learning process, which follows from the study of the quality, robustness, and reliability of a symbolic object.

4.3. Extent of concepts and symbolic objects

When one starts with the description of the extent of a concept, symbolic objects can provide a way to find at least the individuals that constitute this extent.

Example. If the ages of two individuals $w_1, w_2$ are $\text{age}(w_1) = 30$, $\text{age}(w_2) = 35$, then the description of the class $C = \{w_1, w_2\}$ obtained by a generalization process can be $[30, 35]$. The extent of this description contains at least $w_1$ and $w_2$ but may contain other individuals. In this simple case the symbolic object $s$ is defined by a triple: $s = \langle a_C, R_C, d_C \rangle$ where $d_C = [30, 35]$, $R_C = \in$ and $a_C$ is the mapping: $\Omega \rightarrow \{\text{true, false}\}$ such that $a_C(w) = \text{true}$ if $\text{age}(w) \in d_C$. An individual $w$ is in the extent of $s$ iff $a_C(w) = \text{true}$.
More formally (see Figure 2), let $\Omega$ be a set of individuals, $D$ a set containing descriptions of individuals or of a class of individuals, and $y$ a mapping defined from $\Omega$ into $D$ which associates with each $w \in \Omega$ a description $d \in D$ from a given symbolic data table. We denote by $R$ a relation defined on $D$. It is defined by a subset $E$ of $D \times D$. If $(x, y) \in E$, we say that $x$ and $y$ are connected by $R$ and this is denoted by $x R y$. More generally, we say that $x R y$ takes its value in a set $L$. We can have $L = \{\text{true, false}\}$; in this case, $[d' R d] = \text{true}$ means that there is a connection between $d$ and $d'$. We can also have $L = [0, 1]$ if $d$ is more or less connected to $d'$. In this case, $[d' R d]$ can be interpreted as the “true value” of $x R y$ or “the degree to which $d'$ is in relation $R$ with $d$” [see Bandemer and Nather (1992), especially Section 5.2 on fuzzy relations].

For instance, $R \in \{=, \equiv, \leq, \subseteq\}$ or is an implication, a kind of matching that takes missing values, etc. into account. $R$ can also use a logical combination of such operators.

A “symbolic object” is defined by a description $d$, a relation $R$ for comparing $d$ to the description of an individual, and a mapping $a$ called a “membership function.” More formally:

Definition (Symbolic object). A symbolic object is a triple $s = (a, R, d)$ where $R$ is a relation between descriptions, $d$ is a description and $a$ is a mapping defined from $\Omega$ in $L$ depending on $R$ and $d$.

SDA concerns usually classes of symbolic objects where $R$ is fixed, $d$ varies among a finite set of comparable descriptions and $a$ is such that $a(w) = [y(w) R d]$, which is by definition the result of the comparison of the description of the individual $w$ to $d$.

If $y(w)$ and $d$ are trees, we can see that $R$ is a relation between trees.

More generally, many other cases can be considered. For instance, the mapping $a$ is of the following kind: $a(w) = [h_1(y(w)) h_2(R) h_3(d)]$, where the mappings $h_1$, $h_2$ and $h_3$ can be “filters.” There are two kinds of symbolic objects:

(i) “Boolean symbolic objects” if $[y(w) R d] \in L = \{\text{true, false}\}$. In this case, if $y(w) = (y_1, \ldots, y_p)$, the $y_i$ are of types (a) –(d), defined in Section 3.1.

Example (Boolean symbolic object). Let $a(w) = [y(w) R d]$ with $R$: $[d' R d] = V_{i=1,2} [d_i, R_i, d_i]$, where $\vee$ has the standard logical meaning and $R_i = \subseteq$. If $y(w) = \text{color}(w)$, $\text{height}(w))$, $d = \{\text{red, blue, yellow}\}$, $[10,15]) = (d_1, d_2)$, $\text{color}(u) = \{\text{red, yellow}\}$, $\text{height}(u) = \{21\}$, then $a(u) = \text{color}(u) \subseteq \{\text{red, blue, yellow}\} \vee \text{height}(u) \subseteq [10,15]) = \text{true} \vee \text{false} = \text{true}$.

(ii) “Modal symbolic objects” if $[y(w) R d] \in L = [0,1]$. 

Example (Modal symbolic object). Let $a(u) = [y(u) R d]$ where, for instance, $R$: $[d' R d] = \text{Max}_{i=1,2} [d_i', R_i, d_i]$. The choice of the Max is one of many other possible choices related to probabilist, possibilist, and belief objects, and copula theory [see Diday (1995, 1998, 2000a–b), Diday and Vrac, (2005)]. For example, in the case of marginal probability, the likelihood would depend on the product. The relation $R$ between two probability distributions can be defined for two discrete probability distributions $d'_i = r$ and $d_i = q$ of $k$ values by $rRq = \sum_{j=1,k} r_j q_j e^{r_j - \min(r_j, q_j)}$. 

4.4. Syntax of symbolic objects in the case of “assertions”

If the initial data table contains \( p \) variables, we denote \( y(w) = (y_1(w), \ldots, y_p(w)) \), \( D = (D_1, \ldots, D_p) \), \( d \in D: d = (d_1, \ldots, d_p) \) and \( R' = (R_1, \ldots, R_p) \) where \( R_i \) is a relation defined on \( D_i \). We call “assertion” a special case of a symbolic object defined by

\[
s(a, R, d) = \bigwedge_{i=1}^{p} [y_i(w) R_i d_i] \quad \text{where} \quad a \text{ has the standard logical meaning and} \quad a \text{ is defined by} \quad a(w) = [y(w) R d] \text{ in the Boolean case.}
\]

Note that, considering the expression \( a(w) = \bigwedge_{i=1}^{p} [y_i(w) R_i d_i] \), we are able to define the symbolic object \( s = (a, R, d) \). Hence, we can say that this explanatory expression defines a symbolic object called “assertion.”

A sample Boolean assertion is:

\[
a(w) = [\text{age}(w) \subseteq \{12, 20, 28\}] \wedge [\text{SPC}(w) \subseteq \{\text{employee, worker}\}].
\]

If the individual \( u \) is described in the original symbolic data table by \( \text{age}(u) = \{12, 20\} \) and \( \text{SPC}(u) = \{\text{employee}\} \) then \( a(u) = [\{12, 20\} \subseteq \{12, 20, 28\}] \wedge [\{\text{employee}\} \subseteq \{\text{employee, worker}\}] = \text{true}. \)

In the modal case, the variables are multivalued and weighted; an example is \( a(u) = [\text{y}(u) R d] \) with \( [d' R d] = f(\{[y_i(w) R_i d_i]\}_{i=1}^{p}) \) where, for instance, \( f(\{[y_i(w) R_i d_i]\}_{i=1}^{p}) = \Pi_{i=1,2} [d'_i R_i d_i] \).

4.5. Extent of a symbolic object

In the Boolean case, the extent of a symbolic object \( s \), denoted \( \text{Ext}(s) \), is defined by the extent of \( a \), which is \( \text{Extent}(a) = \{w \in \Omega | a(w) = \text{true}\} \). In the modal case, given a threshold \( \alpha \), it is defined by \( \text{Ext}_\alpha(s) = \text{Extent}_\alpha(a) = \{w \in \Omega | a(w) \geq \alpha\} \).

4.6. Concepts: Four approaches

In the Aristotelian tradition, concepts are characterized by logical conjunction of properties. In the Adansonian tradition (Adanson (1727–1806) was a pupil of the French naturalist Buffon, and very much ahead of his time), a concept is characterized by a set of “similar” individuals [Adanson (1757)]. In contrast to the Aristotelian tradition, where all the members of the extent of a concept are equivalent, a third approach, derived from psychology and cognitive science [see Rosch (1978)], is to consider that concepts must be represented by classes which “tend to become defined in terms of prototyped or prototypical instances that contain the attributes most representative of items inside the class” [Rosch (1978, pp. 27–48)]. Wille (1989, pp. 365–380), says “in traditional philosophy things for which their intent describes all the properties valid for the individual of their extent are called ‘concept.’”

Symbolic objects combine the advantages of all four approaches:

(i) The Aristotelian tradition, as they can have the explanatory power of a logical description of the concepts that they represent by “Boolean symbolic objects,” as defined in Section 4.3.

(ii) The Adansonian tradition, as the members of the extent of a symbolic object are similar in the sense that they must satisfy at best the same properties (not necessarily
Boolean, see “modal symbolic objects” in Section 4.3). In that sense, the concepts that they represent are “polythetic.” In this context polythetic (or “polytheistic”) means that such class cannot be described simply by a conjunction of properties.

(iii) The Rosch point of view, as their membership function is able to provide prototypical instances characterized by the most representative attributes. Having a symbolic object $s = (a, R, d_c)$ which models a concept $C$, prototypes of $C$ are instances $w$ which maximize $a(w)$ in the modal case. More details on finding prototypes are given in Section 5.5.


5. Tools for symbolic objects

5.1. Order between symbolic objects

If $r$ is a given order for $D$, then the induced order for the set of symbolic objects denoted by $r_s$ is defined by $s_1 r s_2$ iff $d_1 r d_2$. The result is that, if $R$ is such that $[d R d] = \text{true}$ implies $d' r d'$, then $\text{Ext}(s_i) \subseteq \text{Ext}(s_j)$ if $s_1 r s_2$. If $R$ is such that $[d R d'] = \text{true}$ implies $d' r d$, then $\text{Ext}(s_2) \subseteq \text{Ext}(s_1)$, if $s_1 r s_2$.

5.2. Finding a unique description for a concept: “T-norm of descriptive generalization”

The T-norm operator [Bandemer and Nather (1992), Schweizer and Sklar (1983)] is defined from $[0, 1] \times [0, 1]$ in $[0, 1]$. In order to get a symbolic description $d_c$ of $C$ (i.e., of the concept for which $C$ is an extent), we use an extension to descriptions of the usual T-norm. We call it a “T-norm of descriptive generalization.” This operator is defined on a set of partially ordered descriptions $D$ where $0_D$ and $1_D$ are respectively the smallest and largest descriptions of $D$. For example, if $D$ is the set of intervals included in $[a, b]$ and if the order is the inclusion $\subseteq$, then $1_D = [a, b]$ and $0_D = \emptyset$.

Definition of a T-norm of descriptive generalization. $T$ is a mapping $D \times D \rightarrow D$ satisfying the following four conditions:

(i) $T(u, 1_D) = u$,

(ii) monotonicity: $u1 \leq u2 \Rightarrow T(u1, v) \leq T(u2, v)$,

(iii) commutativity: $T(u, v) = T(v, u)$,

(iv) associativity: $T(u, T(v, w)) = T(T(u, v), w)$.

As $T$ is commutative and associative, we can define a mapping $T^*$ from the power set $P(\Omega)$ (of the set of individuals $\Omega$) into $D$. This mapping is defined as follows: if $A = \{a_1, \ldots, a_n\}$, then $T^*(A) = T(a_n, T(a_{n-1}, T(a_{n-2}, T(\ldots, T(a_2, a_1))))).$ So we have
$T^\ast([a_i, a_j]) = T(a_i, a_j)$. If we denote $y(A) = \{y(w)/w \in A\}$, then $T^\ast(y(A))$ is the description of $A$ for the variable $y$.

In the same way, we can define the T-conorm of descriptive generalization by changing condition (i) $T(u, u) = u$, to (i) $T(u, 0) = u$.

Bandemer and Nather (1992) give many examples of t-norms and t-conorms which can be generalized to t-norms and t-conorns of descriptive generalization. For example, it is easy to see that the supremum and the infimum are respectively a T-norm and a T-conorm. They are also a t-norm and a t-conorm of descriptive generalization. Let $D_C$ be the set of descriptions of the individuals of $C$. The result is that the interval $G(y(C)) = [\inf(D_C), \sup(D_C)]$ constitutes a good generalization of $D_C$, since its extent, as defined by the set of descriptions included in the interval, contains $C$ in a good, narrow way.

Examples of generalization. Let $C = \{w_1, w_2, w_3\}$ and $D_C = \{y(w_1), y(w_2), y(w_3)\} = y(C)$

1. $y$ is a standard numerical variable such that: $y(w_1) = 2.5$, $y(w_2) = 3.6$, $y(w_3) = 7.1$. Then $G(y(C)) = [2.5, 7.1]$ is the generalization of $D_C$ for the variable $y$.

2. $y$ is a probabilistic variable categorical variable, ordered or not ordered (where 1(2/3) means that the probability of the category 1 is 2/3), such that: $y(w_1) = (1(1/3), 2(2/3))$, $y(w_2) = (1(1/2), 2(1/2))$, $y(w_3) = (1(1/4), 2(3/4))$. Then, $G(y(C)) = [(1(1/4), 1(1/2)), [2(1/2), 1(3/4)]$ is the generalization of $D_C$ for the variable $y$.

3. $y$ is a variable whose values are intervals such that: $y(w_1) = [1.5, 3.2]$, $y(w_2) = [3.6, 4]$, $y(w_3) = [7.1, 8.4]$, $C = \{w_1, w_2, w_3\}$. Then, $G(y(C)) = [1.5, 8.4]$ is the generalization of $D_C$ for the variable $y$.

It is not absolutely necessary to describe a class by its t-norm and t-conorm; many alternatives are possible, for instance, if one takes the existence of outliers into account. A good strategy consists of reducing the size of the boundaries in order to reduce the number of outliers. This is done by DB2SO in the SODAS software. Another choice in DB2SO is to use frequencies, in the case of categorical variables.

5.3. Finding several descriptions for a concept

In order to avoid overgeneralization, instead of representing a concept by a unique symbolic description, it may be interesting to obtain a set of symbolic descriptions such that their extents intersect the extent of the concept as well as possible. For example, assume that we need to describe a concept by a set of symbolic objects which satisfies an unsupervised and a supervised criterion simultaneously [see Limam, Diday and Winsberg (2004)]. A top-down clustering tree can be used where at each step a splitting variable is chosen which cuts the extent denoted $A$ of the concept into two subclasses and optimizes a given criterion. This criterion can express (in its unsupervised portion), the sum of the two-by-two dissimilarities of the individuals in each subclass of $A$, and simultaneously (for its supervised portion), the Gini impurity criterion of this class. The process stops when it no longer improves the criterion. In the final tree, we associate with each terminal subclass a symbolic object by the conjunction of the values of the splitting variables used in the branches of the path which defines this subclass. The
extent of each obtained symbolic object partially covers \( A \); together, the set of objects covers \( A \). This method can be applied iteratively to each class of a partition. For example, in the supervised part of the criterion, the variable to be discriminated can be defined by two categories: the given class and its complement.

### 5.4. Dissimilarities between concepts

The dissimilarity between two concepts can be calculated from the dissimilarities (assuming they are known) between the individuals of their extents in a given set \( \Omega \). Hausdorff-type dissimilarity measures are good example of dissimilarity between concepts. Let \( A \) and \( B \) be two subsets of \( \Omega \) for which we know the dissimilarity between the units. Then a Hausdorff-type dissimilarity measure can be defined as follows:

\[
d(A, B) = \max \{ \max_{a \in A} \min_{b \in B} d(a, b), \max_{b \in B} \min_{a \in A} d(a, b) \}.
\]

In the case of several variables, the means of these dissimilarities for each variable can be calculated. The dissimilarity between concepts can also obtained by comparing their symbolic descriptions as in Gowda and Diday (1992) or De Carvalho and Souza (1998). A synthesis and other tools can be found in Bock and Diday (2000).

### 5.5. Finding prototypes from a concept

If we wish to obtain prototypes of a concept from a known extent, we can use a dissimilarity (of the Hausdorff type, for instance) between each member and the other members of this extent. The prototypes are then the members that minimize the sum of these dissimilarities.

When a concept is modeled by a symbolic object, the best prototypes are those that maximize the membership functions of this symbolic object.

If we consider that a prototype is not necessary a member of the set of units, then prototypes can be calculated in various ways depending on the descriptive variable type. For example, if a symbolic variable \( Y \) is of the interval type [i.e., \( Y(w) \) is an interval], a prototype can be defined as follows: let \( M \) be the mean of the means of the intervals associated with the members of the class and \( \Delta \) be the mean of the range of each interval of the class; then the prototype interval can be \( [M - \Delta/2, M + \Delta/2] \). Many other possibilities exist, for instance, using the mean of the min and the mean of the max of each interval of the class. In the case of a variable whose values are subsets of a given set, the probability or the conditional probability (or their product) for each value which appears in the class can be used to define a prototype. In the case of a variable whose values are distributions, their mean or their distribution can be used [see Diday (2002)].

### 6. Underlying structures of symbolic objects

#### 6.1. A generalized conceptual lattice

Under some assumptions concerning the choice of \( R \) and \( T \) (for instance \( T = \max \) if \( R = \leq \) and \( T = \min \) if \( R = \geq \)) it can be shown that the underlying structure of a set of

\( F: \) from \( P(\Omega) \) (the power set of \( \Omega \)) into \( S \) (the set of symbolic objects) such that \( F(C) = s \), where \( s = (a, R, d) \) is defined by \( d = T_{c \in C} y(c) \) and so \( a(w) = [y(w) R T_{c \in C} y(c)] \), for a given \( R \). For example, if \( T_{c \in C} y(c) = \cup_{c \in C} y(c) \), \( R \subseteq \subseteq \), \( y(u) = \{\text{pink, blue}\} \), \( C = \{c, c'\} \), \( y(c) = \{\text{pink, red}\} \), \( y(c') = \{\text{blue, red}\} \), then \( a(u) = [y(u) R T_{c \in C} y(c)] = \{\{\text{pink, blue}\} \subseteq \{\text{pink, red}\} \cup \{\text{blue, red}\}\} \) is true and \( u \in \text{Ext}(s) \).

\( G: \) from \( S \) in \( P(\Omega) \) such that: \( G(s) = \text{Ext}(s) \).

A “complete symbolic object” \( s \) is such that \( F(G(s)) = s \). Such objects can be selected from the Galois lattice but also from a partitioning, a hierarchical or pyramidal clustering, the most influential individuals in a factorial axis, a decision tree, etc.

Example. A symbolic data table is given in Figure 3. With the choice \( T \equiv \text{Max} \) and \( R \equiv \subseteq \), the Galois lattice [Barbut and Monjardet (1971)] obtained from this table is given in Figure 3 as well. This lattice can be constructed, for instance, by using Chein’s (1969) generalization to symbolic data or Ganter’s (1984) algorithm [cited in Diday (1998, 2000), Polaillon (1998)]. The set of all the complete symbolic objects and their extents, which are the vertices of this lattice, are listed below:

\[
\begin{align*}
\text{s}_1: & \quad a_1(w) = [y_1(w) \subseteq O_1] \land [y_2(w) \subseteq O_2] \land [y_3(w) \subseteq O_3], \quad \text{Ext}(s_1) = \{w_1, w_2, w_3, w_4\} \\
\text{s}_2: & \quad a_2(w) = [y_2(w) \subseteq \{e\}] \land [y_3(w) \subseteq \{g,h\}], \quad \text{Ext}(s_2) = \{w_1, w_2, w_4\} \\
\text{s}_3: & \quad a_3(w) = [y_1(w) \subseteq \{c\}], \quad \text{Ext}(s_3) = \{w_2, w_3\} \\
\text{s}_4: & \quad a_4(w) = [y_1(w) \subseteq \{a,b\}] \land [y_2(w) = \emptyset] \land [y_3(w) \subseteq \{g,h\}], \quad \text{Ext}(s_4) = \{w_1, w_2\} \\
\text{s}_5: & \quad a_5(w) = [y_2(w) \subseteq \{e\}] \land [y_3(w) \subseteq \{h\}], \quad \text{Ext}(s_5) = \{w_4\} \\
\text{s}_6: & \quad a_6(w) = [y_1(w) \subseteq \{a,b\}] \land [y_2(w) = \emptyset] \land [y_3(w) \subseteq \{g\}], \quad \text{Ext}(s_6) = \{w_1\} \\
\text{s}_7: & \quad a_7(w) = [y_1(w) = \emptyset] \land [y_2(w) = \emptyset] \land [y_3(w) \subseteq \{g,h\}], \quad \text{Ext}(s_7) = \{w_2\} \\
\text{s}_8: & \quad a_8(w) = [y_1(w) = \emptyset] \land [y_2(w) = \emptyset] \land [y_3(w) = \emptyset], \quad \text{Ext}(s_8) = \{\emptyset\}
\end{align*}
\]

![Fig. 3. The conceptual lattice, the symbolic objects associated with the vertices and their extent, obtained from the given symbolic data table.](image-url)
6.2. Mathematical framework of a symbolic data analysis

Finally, we can summarize the mathematical framework of a symbolic data analysis with the following sets and mappings (see Figure 4): $\Omega$ is the set of individuals, $D$ the description set, $L = \{\text{true, false}\}$ or $L = [0,1]$, $S$ is the set of symbolic objects, $y$ the description function, $a$ the membership function from $\Omega$ in $L = \{\text{true, false}\}$ or $L = [0,1]$, $R$ is the comparison relation, $T$ the generalization mapping, $F$ the intension mapping, $G$ the extension mapping, $d_w$ such that $y(w) = d_w$ is an individual description, $w^s$ is such that $w^s = F(w) = (a, R, y(w))$ is an individual symbolic object, $d_c$ is the description of a class $C$ of individuals, $s$ is the symbolic object given by $F(C) = (a, R, d_c)$ where $a = [y(w) R d_c]$, $G(s) = C$ is the extent of $s$.

7. Steps and tools for Symbolic Data Analysis

7.1. Main steps

In order to apply the tools of standard data mining (clustering, decision trees, factorial analysis, rule extraction, etc.) to concepts, we must extend these tools to symbolic data. We can summarize the main principles of SDAin the following steps:

1. A SDA needs two levels of units. Units at the first level are individuals, and those at the second level are concepts.
2. A concept is described by using the description of a class of individuals defined by the extent of a category.

Fig. 4. Mathematical framework of Symbolic Data Analysis containing the intent and extent mappings $F$ and $G$. 

Ch. 38: Categorization in Symbolic Data Analysis
3. The description of a concept must take account of the variation in the individuals within its extent.

4. A concept can be modeled by a symbolic object which can be improved in a learning process (based on the schema in Figure 2), taking the arrival of new individuals into account.

5. SDA extends standard exploratory data analysis and data mining to cases where the units are concepts described by symbolic data.

6. The output of some SDA methods (clustering, symbolic Galois lattice, decision trees, Kohonen mappings, etc.) provides new symbolic objects associated with new categories (i.e., categories of concepts).

7. Next steps: the new categories can be reified in new concepts as in step 2, and so on.

7.2. Descriptive SDA in SODAS

The main methods of exploratory data analysis have been extended to the case of symbolic data; some of them are indicated in Figure 5. Various authors [Bertrand and Goupil (2000), Billard and Diday (2000, 2002, 2003), Rodriguez (2000)] have made initial efforts to extend “mean,” “correlation,” and “standard deviation” to symbolic data, by studying histograms and covariance of interval data. Several other descriptive methods are available in the SODAS software, such as interactive and ergonomic graphic representations of symbolic descriptions, extension of elementary descriptive statistics to symbolic data (histograms, mean, min, max, biplot, etc., from a symbolic data table), symbolic regression, discrimination, decision trees, clustering (by partition, hierarchy, pyramid), and Principal Component Analysis where the output of this
method preserves the internal variation in the input data in the sense that the individu-
als are not represented in the factorial plane by points as usual but by rectangles. These
rectangles can allow a simpler definition of the same symbolic object with few explana-
tory high-inertia factorial axes since variables, etc., have also been developed.

8. Overview of SODAS

8.1. Some advantages of the use of concepts modeled by symbolic objects

We can observe at least six kinds of advantages resulting from the use of symbolic objects.
1. They summarize the original symbolic data table in an explanatory way (i.e., close
to the initial language of the user) by expressing concepts by means of descriptions
based on properties of the initial variables or meaningful variables (such as indica-
tors obtained by regression or factorial axes).
2. They can be easily transformed in terms of a query in a database and so they can be
used in order to propagate concepts between databases (for instance, from one time
to another time or from one country to another country).
3. Being independent of the initial data table, they are able to identify the matching
with any new individual.
4. If their descriptive portions are used, they are able to generate a new symbolic data
table at a higher level on which a second-level symbolic data analysis can be applied.
5. In order to describe a concept, they are able to easily join several properties based
on different variables coming from different relations in a database and different
population samples.
6. In order to apply exploratory data analysis to several databases, instead of merging
them into one huge database, an alternative is to summarize each one by means of
concepts modeled by symbolic objects and then to apply SDA to the whole set of
obtained symbolic descriptions.

8.2. Overview of SODAS software

In Figure 5 an overview of the SODAS software is given. The input of DB2SO (see
Section 3.4) is a query in a database. Its output is a symbolic data table where the units
are concepts reifying the categories defined by the query. Once this symbolic data table
has been obtained, any of SODAS’s symbolic data analysis methods can be applied.

9. Final remarks

The aim of the EUROSTAT European Community project called SODAS (for Symbolic
Official Data Analysis System), in which 17 institutions in nine European countries are
concerned, was to produce a new methodology and software for SDA. To this end, new
theoretical methods and software developments has been generated. Nevertheless, much still remains open in this vast field of research and development.

We hope that these ideas, based on categories and concepts from computer sciences, will stimulate new directions for research and thinking in the cognitive sciences.

Concerning the free SODAS software:

The SODAS software can be downloaded onto a PC from the website www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm. Much information on the methods and many examples of applications in numerous domains can also be found in this website.

References

Arnault, A., and P. Nicole (1662/1965), La logique ou l’art de penser (Froman, Stuttgart).
Benzecri, J.P. et al. (1973), L’analyse de données (Dunod, Paris).


In this chapter, I describe the implications for a general theory of category formation of a set of experiments in which embodied artificial agents are evolved for the ability to accomplish simple tasks. In particular, I will focus on how categories might emerge from the dynamic interaction between an agent and its environment, and on the relation between categories and behavior. Finally, I will introduce and discuss the notion of action-mediated categories, that is, sensory or internal states that provide indirect and implicit information about the external environment and the agent/environment relationship by exploiting the results of a stereotypical way of interacting with the environment.
1. Introduction

A new research paradigm, which has been called *Embodied Cognitive Science* [Pfeifer and Scheier (1999)], has recently challenged the traditional view that intelligence is an abstract process that can be studied without taking into consideration the physical aspects of natural systems. In this new paradigm, researchers tend to stress *situatedness*, i.e., the importance of studying systems that are situated in an environment [Clark (1997), Brooks (1991)]; *embodiment*, i.e., the importance of studying systems that have bodies, receive input from their sensors, and produce motor actions as output [Brooks (1991), Clark (1997)]; and *emergence*, i.e., the importance of viewing behavior and intelligence as the emergent result of fine-grained interactions between an agent’s control system including its constituent parts, its body structure, and the external environment. An importance consequence of this view is that the agent and the environment constitute a single system, i.e., the two aspects are so intimately connected that a description of either one of them in isolation does not make much sense [Maturana and Varela (1980, 1988), Beer (1995)].

An attractive way of studying embodied and situated agents consists in developing these systems through a self-organizing process, such as artificial evolution, that allows them to develop their skills autonomously in close interaction with the environment and without human intervention [Nolfi and Floreano (2000)]. Recent experimental research in this area has shown how self-organizing artificial agents might develop simple cognitive abilities such as the ability to integrate sensorimotor information over time and form internal categories [Nolfi and Tani (1999), Slocum, Downey and Beer (2000), Nolfi and Marocco (2001), Beer (2003)].

In this chapter, I will review these recent findings and their implications from the point of view of category formation. Rather than focusing on how a shared language can self-organize in a population of interacting embodied agents (in this issue, see Cangelosi’s chapter in this book), I will focus on how categories might emerge from the dynamic interaction between an agent and its environment, and on the relation between categories and behavior. In particular, I will introduce and discuss the notion of *action-mediated* states, that is, sensory or internal states that provide indirect and implicit information about the external environment and the agent/environment relationship by exploiting the results of stereotypical ways of interacting with the environment.

2. The method

One effective way to build artificial agents that are able to develop their skills autonomously in close interaction with the environment is to rely on evolutionary computation and, more specifically, on evolutionary robotic techniques [Nolfi and Floreano (2000)].

The basic idea behind this approach is as follows: an initial population of different artificial genotypes, each of which encodes the control system (and sometimes the morphology) of a robot, is created randomly. Each robot (physical or simulated) is
placed in the environment and left free to act (move, look around, manipulate) while its performance on various tasks is automatically evaluated. The fittest robots are allowed to reproduce by generating copies of their genotypes with the addition of changes introduced by some genetic operators (e.g., mutations, crossover, duplication). This process is repeated for a number of generations until an individual is born that satisfies the performance criterion (fitness function) set by the experimenter.

The experimenter must design the fitness function, which is a criterion that is used to measure how well an individual robot is able to accomplish the desired task. Moreover, the experimenter must specify how genetic information (which is usually encoded as a sequence of binary values) is expressed in the corresponding phenotypical robot. However, the mapping between genotype and phenotype is usually task-independent, and evolving individuals are selected only on the basis of the overall efficacy of their behavior. This allows to minimize our a priori ideas of how a given problem should be solved; thus the robots are free to explore the array of possibilities in a relatively unbiased way.

3. Categories emerging from the interaction between the agent and the environment

In this section, we will show how problems that apparently require agents to be able to discriminate among different categories (i.e., different classes of environmental situations) can be solved by relying on simple control strategies that do not require the agent to internally assign environmental situations to separate classes eliciting different motor responses.

3.1. Finding and remaining in favorable environmental areas

Consider the case of a simulated agent that lives inside a circular strip divided into 40 cells (20 cells on the left and 20 on the right). At each point in time, the agent occupies a single cell and perceives a sensory state corresponding to the cell type. There are 20 different cell types and 20 different sensory states that the agent can perceive, numbered from 0 to 19. Cell types are distributed in a randomly generated fashion, but each cell type is present once in the left and once in the right side of the environment (see Figure 1, left). The agent can react to its current sensory state in two different ways (move one cell away, in either a clockwise or counterclockwise direction). The agent’s goal is to reach and remain in the left side of the environment [Nolfi (2002b)].

Agents have a neural network with 20 sensory neurons that locally encode the corresponding perceived sensory state and one output unit that binarily encodes one of the two possible actions (see Figure 1, right). As a consequence, only one sensory neuron is activated at each point in time. Weights can assume only two values (−1 or 1). Consequently, the weight of the connection between the current activated sensory neuron and the motor neuron locally encodes how the agent reacts to its current sensory state (i.e., the agent moves clockwise or counterclockwise when the connection weight is −1 or 1, respectively). Agents do not have any memory of the sensory states they have previously experienced (i.e., they always react in the same way to a given sensory state).
What is interesting about this experimental situation is that none of the possible sensory states provides in itself any indication of the agent’s current location. For each possible sensory state, in fact, agents have a 50% probability of being in the left or in the right part of the environment. Apparently, therefore, agents that only decide how to act on the basis of the current sensory state cannot solve this problem.

However, by evolving agents for the ability to move toward the left side of the environment\(^1\), we observed that, after a few generations, evolving individuals were able to move away from the right side and to remain in the left-hand half of the environment. The way in which evolved individuals solve this problem can be seen by observing the arrows on the right-hand half of the circular environment shown in Figure 1. On the right side of the environment, individuals consistently move clockwise or counterclockwise until they exit from the right side. Conversely, in some areas of the left side of the environment, individuals start to move back and forth, remaining there for the rest of the epoch.

---

\(^1\) Evolving individuals were allowed to “live” for 100 epochs, with each epoch consisting of 200 actions. At the beginning of each epoch, agents are placed in a randomly selected location of the environment. Fitness is computed by counting the number of epochs in which individuals were located on the left-hand portion of the environment after 200 cycles. Connection weights were binarily encoded in the genotype, which was 20 bits long. Population size was 100. The best 20 individuals in each generation were allowed to reproduce by generating five copies of their genotype with 2% of their bits replaced by a new, randomly selected value.
Individuals always react to the same sensory state in the same way. Despite that, the way in which they react to the different sensory states allows them to produce behavioral attractors in the left but not in the right part of the environment. Attractors consist of pairs of adjacent cells to which the agent reacts by moving clockwise and then counterclockwise (when moving in a clockwise direction, the robot should respond clockwise to the first cell and counterclockwise to the second cell – see the points indicated by an $a$ in Figure 1, left). When the agent encounters an attractor point, it remains there by moving back and forth repeatedly. For an example in which the same type of strategy emerges by evolving a Khepera robot for the ability to discriminate between objects with different shapes, see Nolfi (2002b).

3.2. Discriminating objects with different shapes on the basis of tactile information

As second example, consider the case of a robot with an artificial finger that has to discriminate between objects with different shapes on the basis of rather rough tactile information [Nolfi and Marocco (2002)].

The finger consists of three segments with six degrees of freedom (DOF) and extremely coarse touch sensors (see Figure 2, left). More specifically, the artificial finger has a basic structure of two bodies and two joints replicated three times (see Figure 2, right). These two bodies are connected by means of a joint (i.e., Joint $E$ in Figure 2, right) that allows only one DOF on the $Y$–axis, while the shorter body is connected to the floor or to the longer body, by means of another joint (i.e., Joint $R$) that provides one DOF on the $X$–axis. In practice, Joint $E$ allows the agent to raise and lower the connected segments and Joint $R$ allows it to rotate them in both directions. Joint $E$ and Joint $R$ are free to move only in a range between $[0, \pi/2]$ and $[–\pi/2, +\pi/2]$, respectively. Each actuator is provided with a corresponding motor that can apply a varying force.

![Fig. 2. Left: the artificial finger and a spherical object. Right: a schematic representation of the finger.](image-url)
Therefore, to reach every position in the environment, the control system must appropriately control several joints and deal with the constraints due to gravity.

The sensory system consists of three simple contact sensors placed on each longer body that detect when it collides with obstacles or other bodies and six proprioceptive sensors that indicate the current position of each joint. The controller of each individual consists of a neural network with 10 sensory neurons directly connected to seven motor neurons and two internal neurons that receive connections from the sensory neurons and from themselves and project connections to the motor neurons. The first nine sensory neurons encode the angular position (normalized between 0.0 and 1.0) of the six DOF of the joints and the state of the three contact sensors located in the three corresponding segments of the finger. The last sensory neuron is a copy of the last motor neuron that encodes the current classification produced by the individual (see below). The first six motor neurons control the actuators of the six corresponding joints. The output of the neurons is normalized between $[0, +\pi/2]$ and $[-\pi/2, +\pi/2]$ in the case of elevation and rotational joints respectively and is used to encode the desired position of the corresponding joint. The motor is activated so as to apply a force proportional to the difference between the joint’s current and desired positions. The seventh motor neuron encodes the classification of the object produced by the individual (values below or above 0.5 are interpreted as classifications corresponding to a cubic or spherical object respectively).

By running 10 replications of the experiment and by evolving individuals for 50 generations, we observed that in many of the replications evolved individuals displayed a good ability to discriminate between the two objects and, in some cases, they produced close to optimal performance.

By analyzing the obtained behaviors, one can clearly see that in all experiments evolved individuals select a well-defined behavior that assures that perceived sensory states corresponding to different objects can be easily discriminated and allows robust and effective categorizations. Figure 3 shows how a typical evolved individual behaves with a spherical object.

---

2 Evolving individuals were allowed to “live” for 36 epochs, each epoch consisting of 150 actions. Each individual of the population was tested for 36 epochs, with each epoch consisting of 150 life cycles. At the beginning of each epoch, the finger was fully extended and a spherical or cubic object was placed in a randomly selected position in front of the finger (the position of the object was randomly selected from the following intervals: $20.0 \approx X \approx 30.0; \ 7.5 \approx Y \approx 17.5; \ -10.0 \approx Z \approx 10.0$). The object is a sphere (15 units in diameter) in even epochs and a cube (15 units along the side) in odd epochs, so that each individual had to discriminate the same number of spherical and cubic objects during its “lifetime.” Fitness was computed by counting the number of epochs in which individuals correctly classified the object (i.e., the number of times in which, at the end of the epoch, the activation of the last motor unit was below 0.5 if the object was a cube or above 0.5 if the object was a sphere). Population size was 100. The best 20 individuals in each generation were allowed to reproduce by generating five copies of their genotype with 1% of their bits replaced by a new, randomly selected value.
and a cubic object (left and right sides of the figure, respectively). As can be seen, first the finger bends on the left side and moves to the right in order to start to feel the object with the touch sensor on the third segment. Then the finger moves so as to follow the curvilinear surface of the sphere or to keep touching one of the corners of the cube.

The fact that such behavior significantly simplifies the discrimination task can be explained by considering that the finger ends up in very different conditions in the case of a sphere or of a cube. In particular, after a certain amount of time in which the finger is negotiating with the object, it ends up almost fully extended in the case of spheres and almost fully bent in the case of cubes. This implies that, in light of such behavior, the state of the proprioceptive sensors after a certain amount of time can be used as a direct and straightforward indication of the category of the object. The fact that such behavior allows evolved individuals to effectively discriminate between the two objects can be explained by considering that the discrimination process is not the result of a single decision but the end result of an interaction between the agent and the object that lasts for several cycles. Indeed, evolved individuals that display shorter negotiation periods with spherical objects also perform worse (results not shown). A similar temporally extended decision process was observed by Beer (2003) in evolved agents asked to catch diamond-shaped objects and avoid circular objects.

3.3. Behavior emerging from the dynamic interaction between the agent and its environment

The two examples reported in the previous sections show how the ability to categorize objects or environmental situations does not necessarily require agents to be able to assign sensory states or sequences of sensory states to different internal categories.

To understand this apparent paradox we should distinguish between two ways of describing behavior. A distal description of behavior is a description from the observer’s point of view, in which high-level terms such as “approach” or “discriminate” are used to describe the result of a sequence of sensorimotor loops. A proximal description of behavior is a description from the point of view of the agent’s sensorimotor system that describes how the agent reacts to different sensory and internal states (see Figure 4).

It should be noted that behavior from the point of view of a distal description is the result not only of behavior from a proximal point of view but also of the environment. More specifically, it is the result of the dynamic interaction between the agent and the environment. The sensory patterns that the environment provides to the agent partially determine the agent’s motor reactions. These motor reactions, in turn, by modifying the environment or the relative position of the agent in the environment, partially determine the types of sensory patterns that the agent will receive from the environment.

The fact that behavior, from a distal perspective, emerges from the dynamic interaction between the agent and the environment implies that there is not necessarily an one-to-one correspondence between the distal and proximal descriptions of behavior, and there is no reason to expect that what makes sense at the distal level of behavior will also makes sense at the proximal level.
The fact that an agent is able to discriminate between different types of objects or different environmental situations (by producing different labels for them or by reacting differently in different situations) from the point of view of a distal description of behavior, therefore, does not necessarily imply that a discrimination process is occurring in the agent’s control system (i.e., at the level of the proximal description of behavior).

Indeed, the evolved agent described in Section 3.1 never “knows” whether it is located in the good or bad side of its environment and reacts in the same way to both sides of the environment. It simply acts in such a way that, by interacting with that environment, it will always leave the bad side and remain in the good side. Similarly, the evolved agent described in Section 3.2 does not “know” whether the object it is currently touching is a cube or a sphere. It simply acts in a way that, by interacting with the object, produces two qualitatively different behaviors in the two cases (i.e., to keep touching the cube and to stop touching the sphere).

These two examples are a straightforward demonstration of the importance of embodiment, situatedness, and emergence. Regarding embodiment, we should consider
that, for example, the behavior of the artificial finger and its ability to discriminate different shapes strongly depend on the physical shape of the finger itself and on the results of its physical collisions with the objects. Regarding situatedness, we must remember that the dynamic interaction with the environment and the structure of the environment play a crucial role in the way in which the problem is solved. In the case of the agent living in a circular environment, for instance, the ability to produce dynamic movement (e.g., moving back and forth in the attractor areas of the environment) clearly results from the dynamic interaction between the agent’s control system and its environment. Finally, regarding emergence, we should bear in mind the fact that evolved solutions are typically qualitatively different from the solutions that we, as external designers, tend to develop. For a human observer, solving the problem of the circular environment with a simple solution qualitatively similar to those discovered by artificial evolution is, in fact, extremely hard, if not impossible.

4. Action-mediated sensory states

The case of the artificial finger described in Section 3.2 is also interesting for another aspect that we will discuss in more detail in this section. The well-defined way with which evolved individuals interact with their environment not only allows them to display two different behaviors in the case of the two categories (i.e., cubes and spheres). It also ensures that, after a certain amount of time, the activation state of the sensors is well differentiated for different types of objects. After a short interaction with the objects, in fact, the finger is bent in the case of cubes and extended in the case of spheres. After this short negotiation phase, therefore, the activation state of the proprioceptive sensors that encode the positions of the six actuators becomes well differentiated in the two cases. This explains how the artificial finger is able to appropriately label the two objects.

More generally, we can say that well-defined ways of interacting with the environment might allow sensory states to indirectly convey complex information about the external environment that would not become available without such interaction. These states are action-mediated, given that they do not convey such information by themselves; rather, they only acquire their meaning after an appropriate interaction with the environment has taken place. For example, the state of the proprioceptive sensors corresponding to a bent finger do not provide any information about the shape of the object placed close to the finger; they only provide this information if the finger has previously interacted with that object on the basis of the simple behavior described above.

In this section, we will describe two other experiments in which evolved agents are able to solve their adaptive tasks by selecting simple ways of interacting with the environment that, in turn, ensure that they will later experience useful action-mediated sensory states (i.e., sensory states providing ready-to-use information for discriminating between different types of objects or different environmental locations).
4.1. Discriminating larger and smaller cylindrical objects

Consider the case of a mobile robot placed in an environment surrounded by walls; this robot is required to find and stay close to large cylindrical objects and avoid small cylindrical objects. The robot is a Khepera [Mondada, Franzi and Ienne (1993)], a miniature mobile robot with a diameter of 55 mm and a weight of 70 g. It is supported by two lateral wheels that can rotate in both directions and two rigid pivots in the front and in the back. The sensory system employs eight infrared sensors that are able to detect obstacles up to about 4 cm away.

As demonstrated by Scheier, Pfeifer and Kuniyoshi (1998) and Nolfi (2002b), this problem is far from trivial given that the categories corresponding to large and small objects largely overlap in sensory space. Indeed, the distance in sensory space for sensory patterns originating from a single object can be large, while the distance in sensory space for sensory patterns originating from two different objects can be small. Nevertheless, evolved robots are able to solve the problem on the basis of a simple control strategy [Scheier et al. (1998)].

As in the case of the experiments reported in Section 3, the authors evolved the connection weights of the robots’ neural controllers³. By analyzing the performance of robots through out generations, Scheier et al. (1998) observed that it improves rather quickly during the first generations and stabilize around near optimal performance after about 40 generations. The fittest individuals in 86% of the runs move in the environment until they start to perceive an object (large or small) and then start to circle around the object. (The individuals in the other 14% of runs stop in front of the objects; however, these individuals perform significantly worse.) The robots continue to circle around large objects while avoiding and abandoning small objects. This circling behavior is crucial to discriminating between the two types of objects. In fact, while the sensory patterns corresponding to small and large cylinders largely overlap overall, the subsets of sensory patterns experienced while the robot is circling small and large cylinders are nicely separated in sensory space. In other words, the circling behavior allows the robot to select sensory patterns that can be easily discriminated.

The importance of the circling behavior can be further demonstrated by analyzing the complexity of the discrimination task for individuals of successive generations. To understand this point, we should consider that, given that the type of sensory patterns that an individual receives from the environment partially depends on how that individual reacts

³ Evolving individuals were allowed to “live” for five epochs, with each epoch consisting of 5000 actions. Individuals’ fitness was increased each time they were close to a large object and decreased when they were close to a small object or a wall. Connections were represented in the genotype by a six-bit string, where one bit determined whether the connection was to be used or not and five bits coded for the strength of the corresponding weight. Population size was 50. The 10 best individuals in each generation were allowed to reproduce by generating five copies of their genotype, which were mutated by replacing 5% of randomly selected bits by a new, randomly chosen value. The experiment was replicated 30 times using four different network architectures (with and without recurrent connections, and with and without hidden units). Similar results were obtained for all types of architecture.
to each sensory state, individuals that differ in their way of interacting with the environment may face simpler or harder discrimination tasks.

To quantify the complexity of the discrimination task, the authors measured how separate the two classes of sensory patterns corresponding to the two kinds of objects (small and large cylinders) were in the sensory space. This measure can be obtained by using the geometric separability index (GSI) proposed by Thornton (1997), which provides a quantitative measure of the separation in space of two or more classes of sensory patterns.

In the case of these experiments, GSI can be computed by storing all sensory patterns experienced by an individual during $N$ life cycles and by checking, for each sensory pattern, whether the nearest pattern (Euclidean distance) belongs to the same class. The total number is then normalized by $N$. If the nearest pattern in sensory space always belongs to the same class as the currently perceived object, the GSI value is 1: this means the patterns of the two categories are well separated. Values close to 1 therefore indicate that the sensory patterns belonging to the two categories are quite separated in the input space and easy to discriminate, while values close to 0.5 indicate that the sensory patterns corresponding to the two categories largely overlap (see Figure 5).

As reported by Scheier et al. (1998), the GSI value of the best individuals in successive generations starts at about 0.5 and monotonically increases during the first 40 generations until it reaches a stable state at around 0.9 (note that performance also improves during the first 40 generations). This means that the ability of evolving individuals to find and stay close to large cylinders while avoiding small cylinders is mainly due to an ability to act in such a way as to experience sensory patterns that can be easily discriminated.

Fig. 5. A schematic representation of the distribution of sensory patterns. For the sake of simplicity, the sensory space has only two dimensions (S1 and S2). Crosses and circles represent all possible sensory patterns originating from small and large cylinders, respectively. Dark crosses and circles represent the sensory patterns experienced by a given individual. Note that individuals tend to experience only a subset of all possible sensory patterns and that the patterns forming the subset depend on the type of behavior displayed by that individual. The three figures indicate the sensory patterns experienced by three different individuals. As shown in the figure, the geometric separability index (GSI) may vary from 0.5 to 1.0, depending on how much the two groups of patterns overlap in sensory space.
Sensory patterns experienced by evolved robots showing the circling behavior, therefore, are another example of action-mediated sensory states. The sensory states experienced while the robot is circling around one large cylindrical object, for example, cannot be separated from all the other possible sensory states. However, they can be easily separated from the sensory states that the robot experience in other situations, if that robot interacts with the environment in a well-specified way.

By evolving agents to visually discriminate between circular and diamond-shaped objects and to catch the former and avoid the latter, Beer (2003) observed that evolved agents foveated and actively scanned any object before eventually catching or avoiding it. According to Beer, the scanning behavior might have the same functional role as the circling behavior described above. Indeed, Beer (2003; p. 214) claims “… it is likely that this scanning accentuates the small differences between a circle and a diamond.”

4.2. Navigating toward a target area of the environment

As a second example, consider the case of a Khepera robot placed in a randomly selected location in a rectangular environment; the robot is supposed to navigate toward the northwest or southeast corners of the environment (Figure 6). The size of the environment and the proportions of the long and short walls vary randomly in each trial within a given range.

Apparently, the only way to solve this problem is to discriminate between long and short walls and to navigate accordingly. For example, the robot could follow a long wall on its own right side until it reaches a corner (or it could use other similar strategies such as following a short wall on its own left side until it comes to a corner). Given that sensors can only provide information about the local portion of the environment surrounding the robot (i.e., they are activated by obstacles up to 4 cm away) and given that the size of the environment may vary, the ability to detect long or short walls seems to require an ability to: (1) “measure” the length of two adjacent walls by moving along them and identifying the beginning and the end of each wall, (2) “memorize” the measured lengths in

![Fig. 6. The environment and the robot. The lines represent the walls. Full and dashed arcs represent the right and wrong areas, respectively. The small circle represents the Khepera robot.](image)
internal states of the robot’s controller, and (3) “compare” the two measured lengths stored in internal states.

By selecting robots for the ability to reach the two target areas, however, Nolfi and Marocco (2002) observed that evolving robots provided with simple reactive neural controllers can solve this problem quite well (up to 85% of correct navigations in the case of the best replications of the experiment). These simple neural controllers do not have any internal states and therefore cannot perform the complex measuring behavior described above.

Figure 7 shows a typical strategy displayed by evolved individuals. The robot explores the environment by avoiding walls and by moving forward in straight lines until it finds one of the corners (in the case shown in the figure, the robot starts from the right side of the environment and encounters the bottom left corner first). When it is facing a corner, the robot moves left and right and back and forth for a while and then leaves the corner at an angle of about 45° to the two walls that form the corner. Finally, when it encounters a wall, it turns counterclockwise until the wall is on its right side, and then follows the wall until it reaches a corner.

Fig. 7. A typical strategy displayed by evolved individuals. The line within the rectangle represents the trajectory of a typical evolved individual during a trial.

4 The architecture of the neural controller was fixed and consisted of a fully connected perceptron with eight sensory and two motor neurons encoding the state of the robot’s eight infrared sensors and the speed of the two motors controlling the two wheels. Individuals’ fitness was increased or decreased by one point each time, they ended their lifetime in one of the right or wrong corners, respectively. The genotype encoded the connection weights and biases of the neural controller. Each weight was represented in the genotype by an eight-bit string and normalized to between $-10$ and $+10$. Population size was 100. The best 20 individuals in each generation were allowed to reproduce by generating five copies of their genotype with 4% of their bits replaced by a new, randomly selected value. Individuals were tested for 10 epochs. At the beginning of each epoch, the length of the short–and long walls was randomly selected within $[15, 45]$ and $[16, 90]$ cm, respectively, and the robot was placed in a randomly selected location within the environment with a randomly selected orientation.
This strategy ensures that, after the first corner (which might be any corner given that the initial position and orientation of the robot is randomly chosen), the robot will always reach one of the two long walls. At this point, it can easily find the target area by turning until the wall is on its right side, and then following it. Note that this strategy works for any rectangular environment independently of the relative length of long versus short walls and of the size of the environment. Indeed, leaving corners at an angle of 45° is a “smart” way of measuring the relative length of the walls. Once again, action mediation (i.e., leaving corners at an angle of 45° in the case of this experiment) allows sensory states experienced later on to acquire useful meanings (i.e., sensory states corresponding to walls uniquely identify long walls and sensory states corresponding to corners uniquely identify the two target corners).

This simple strategy, however, does not allow evolving individuals to remain in the target corners. They spend some time there, moving back and forth, but they eventually abandon the corner they are in, quickly moving toward the other correct corner. Indeed, this is why they do not achieve optimal performance (fitness is computed by looking at how many trials end with the robot in one of the two target corners). This inability to remain in the target corners can be explained by the hypothesis that evolved robots “know” how to move to reach the two target corners but do not “know” whether the corner in which they are currently located is correct or not.

Further experiments conducted on evolving robots provided with a neural network with internal neurons and recurrent connections show how, in this case, evolving individuals are also able to remain in one of the two target corners after abandoning one of the other two corners [Nolfi and Marocco (2002)]. Interestingly, the analysis of how evolved individuals solve the problem of finding and remaining in one of the two correct corners indicates that the same strategy described above is used to reach the correct target corners. Internal neurons only keep track of how much time has passed since the robot started to interact with the environment. If enough time has passed and the robot is in a corner, it stops there. This simple behavior exploits the fact that leaving corners at an angle of about 45° and then following walls by keeping them on the right side guarantees that eventually the robot will only encounter target corners.

This example also shows how suboptimal simple strategies based on action-mediated sensory states might be complemented with additional simple internal mechanisms. This possibility is important from an evolutionary or developmental perspective. In fact, it implies that simple strategies based on action-mediated sensory states might later be enhanced in an incremental fashion without necessarily undergoing profound reorganizations [Nolfi and Marocco (2002)].

5. Integrating sensorimotor information over time and the emergence of complex internal categories

The examples described in Sections 3 and 4 show how nontrivial problems can be solved without relying on internal categories but rather by exploiting action-mediated
sensory states that provide the information necessary to behave correctly, at the right time and in a ready-to-use fashion. At this point, we might be interested in trying to understand in what conditions embodied agents might be unable to solve their adaptive problems by relying only on simple reactive or quasireactive solutions.

From this perspective, interesting candidate situations are those in which agents cannot freely select how they will interact with the environment. Limiting their interactions, in fact, by reducing the chances that useful properties emerging from these interactions can be exploited, might prevent agents from exploiting action-mediated sensory states. In these cases, more complex strategies based on internal states might be the only options available.

Several factors might prevent agents from freely determining how they interact with the environment. One case is situations, where the environmental structure strongly limits the freedom of the agent’s behavior. A second class of cases might be constituted by situations in which the agent’s behavior must satisfy several constraints at the same time. Finally, another class of cases might be constituted by agents that are required to be able to communicate their sensorimotor experience to other agents when requested. A case in which the interaction between the agent and the environment is limited by both the environmental structure and the need to communicate will be reviewed in the next section.

5.1. The self-localization problem

Consider the case of a Khepera robot that must travel along a looping corridor (see Figure 8, left) and self-localize by identifying its current location in the environment [Nolfi (2002a)].

Each individual’s controller consists of a neural network with nine sensory neurons directly connected to three motor neurons and five internal neurons that receive connections from the sensory neurons and send connections to the motor neurons and to themselves (see Figure 9). The first three sensory neurons encode the state of the three corresponding motor neurons at the previous point in time; the other six sensory neurons encode the six frontal infrared sensors (normalized between [0.0, 1.0]). The first two motor neurons encode the desired speed of the two corresponding wheels and the last motor neuron encodes the robot’s self-localization output (see below). Internal neurons were updated according to an activation function with a genetically encoded time-constant parameter (which allows neurons to change their activation states at different rates) and a thresholded activation function [see Nolfi (2002a) for details].

The fitness function has two components that reward, respectively, the ability to travel clockwise along the corridor and the ability to indicate the robot’s current position in the environment [see Nolfi (2002a) for details].

Evolved individuals show close to optimal performance on both the navigation and localization tasks. Figure 10 displays the behavior and neural activity of one evolved individual. As shown in the figure, the internal neuron i1 is switched off when the robot negotiates right turns and increases its activation when the robot travels along straight corridors. Thanks to a recurrent positive connection, however, the neuron is switched
off on right turns only if its activation level is below a given threshold (on left corners, on the other hand, this neuron is switched off independently of its activation state – see the point indicated with C). The final result is that this neuron is always below a given

Fig. 8. Left: the environment consists of a looping corridor $40 \times 70$ cm. Lines represent walls. Circles represent cylindrical objects. Arrows represent the starting positions and orientations in which the robot is placed at the beginning of each trial. Center: the environment is divided into 22 idealized regions placed along the corridor clockwise from the top left corner. Right: the environment is ideally divided into two rooms indicated with light and dark gray colors.

Fig. 9. The architecture of the neural controller.
threshold in the light gray room due to the resetting of its activity that occurs in C and A corners but is always over that threshold in the dark gray room. Note that internal neuron i1 captures sensorimotor regularities that extend over rather long time scales (ranging from a few to several seconds). Indeed, in order to self-localize, this robot is able to detect regularities such as corners (which extend over a period of few hundreds of milliseconds) and regularities such as corridors of different lengths (which extend from a few to several seconds). [Note: not all right turns are indicated with an A, and not all As indicate right turns (one marks a left turn, assuming the robot is moving clockwise)].

Note how, contrary to the experiments described in Sections 3 and 4, in the experiment described here evolved robots develop internal categories, that is, internal states that integrate sensorimotor information through time and covary with the current position of the robot in the environment. In the case of the evolved individual described in Figure 10, for example, the internal neuron i1 encodes the distance traveled by the robot since the last left turn or the last right turn followed by a short corridor. In other evolved individuals, internal neurons simply encode the distance traveled from the last left turn [see Nolfi (2002a)], or the frequency with which the robot encountered left and right turns during its previous movements, weighted by the type of turn [see De Croon, Nolfi and Postma (in press)].
The possibility of relying on simpler strategies exploiting action-mediated sensory states is prevented by the need to move fast and by the structure of the environment which, being made up of tight corridors, leaves the robot very little freedom concerning how to move about in the environment. Indeed, in experiments in which agents were unable to extract relatively “complex” internal categories like those described above, we observed that evolving individuals: (1) performed very poorly on the self-localization task when asked to travel fast (i.e., when asked to visit at least 1000 areas of the environment in succession), and (2) performed well but suboptimally when allowed to travel at a lower speed [De Croon et al. (in press)]. In the latter case, evolved individuals exploited action-mediated sensory states that allowed them to partially solve the self-localization process without relying on internal categories. For example, some evolved individuals traveled along corridors by moving slightly from the left to the right side of the corridor. This allowed them to experience a unique sensory state toward the end of the long corridor which, in turn, enabled them to detect the beginning of the dark gray room without internally encoding the lengths of corridors.

6. Conclusions

In this chapter, I described the results of a series of experiments in which embodied artificial agents autonomously develop their abilities in interaction with the environment, thanks to a self-organizing process based on artificial evolution.

By analyzing the evolved individuals we observed that, by exploiting properties that emerge from well-specified ways of interacting with the environment, they can solve nontrivial problems without the need to develop internal categories and, more generally, without the need to internally process sensorimotor information. By selecting well-defined ways of interacting with the environment, in fact, evolved individuals are able to experience action-mediated sensory states that provide ready-to-use information when needed (i.e., information that can be transformed directly into the appropriate motor actions without significant further elaboration).

We also showed how simple reactive strategies based on the exploitation of action-mediated sensory states might be complemented with an ability to integrate sensorimotor information over time into internal states that can later be used to appropriately modulate the agent’s behavior.

Finally, we showed how the need to rely on internal categorization and, more generally, on the internal elaboration of sensorimotor information tends to be particularly compelling in the case of agents that, due to environmental and adaptive constraints, are unable to freely choose between different ways of interacting with the environment.

The results and the analysis reported in this chapter demonstrate that the evolutionary method is a powerful tool for understanding adaptive behavior in embodied and situated agents. It provides a way of understanding how behavior emerges from the interaction between the control system, the body, and the environment.
Acknowledgments

This research was partially supported by the ECAGENTS project founded by the European Community Future and Emerging Technologies program (IST-FET) under EU R&D Contract 001940. The information provided is the sole responsibility of the author and does not reflect the Community’s opinion. The Community is not responsible for any use that may be made of data appearing in this publication.

References


AN INFORMATION-BASED DISCUSSION OF VAGUENESS*: SIX SCENARIOS LEADING TO VAGUENESS

DIDIER DUBOIS AND HENRI PRADE
IRIT CNRS, Université Paul Sabatier, 31062 Toulouse Cedex 9, France

FRANCESC ESTEVA AND LLUIS GODO
IIIA CSIC, Campus Univ. Aut. Barcelona, 08193 Bellaterra, Spain

Contents

Abstract 892
1. Introduction 893
2. The information framework 894
3. Classical vs. gradual properties 895
   3.1. Graduality and partial preordering 895
   3.2. Membership functions as total preorders 896
   3.3. Fuzzy sets and similarity to prototypes 897
   3.4. Set-theoretic operations 897
   3.5. Graduality is a useful form of vagueness 898
4. Precisely defined vs. poorly defined properties 899
   4.1. Classification ambiguity 900
   4.2. Vagueness as limited perception 900
   4.3. Supervaluations 901
   4.4. Ill-known partial membership 901
5. Refining precisely defined properties using closeness relations 901
6. Single agent vs. multiple agents 902
7. Ill-known attribute values and twofold sets 904
8. Approximately described sets 905
9. Concluding remarks 906
References 907

* This chapter is a fully revised and expanded version of a conference paper with the same title, presented at the 10th IEEE International Conference on Fuzzy Systems, Melbourne, Australia, December 2–5, 2001.
Abstract

The issue of understanding and modeling vagueness has been addressed by many authors, especially in the second half of the twentieth century. In this chapter, we try to provide an organized discussion of different categories of vagueness, pointing out circumstances, where they appear. Together, they lead to a trichotomy of the universe of discourse, which seems to be the common feature of the different forms of vagueness. Basic representational frameworks are proposed for each case. This chapter does not advocate a particular view but identifies the characteristic features of each situation.
1. Introduction

Vagueness has been discussed for a long time by philosophers and logicians [Pierce (1878, 1931), Russell (1923), Hempel (1939), Fine (1975), Machina (1976), Williamson (1994), Haack (1996), Keefe (2000), Sorensen (2001)]. It is generally considered in relation to the “sorites paradox,” and the failure of the principle of bivalence in logic. As pointed out in Sanford (1995), there are two drastically opposite approaches to vagueness: supervaluations (preserving a form of bivalence but admitting truth-value gaps), and degrees of truth (rejecting the bivalence principle). In the first view [Van Fraassen (1969)], predicates are Boolean, but their extension may be unknown. In the second view, some predicates are intrinsically non-Boolean. The idea that truth is a matter of degree is already advocated in the philosophical works of Bradley (1914, pp. 257–258) (“All truths and all errors in my view may be called relative, and the difference in the end between them is a matter of degree”).

Vagueness is usually viewed as a defect. However, Black (1937) did not see vagueness as necessarily constituting a defect, and distinguished it from both generality (or nonspecificity) and ambiguity (e.g., a word with several interpretations). He first proposed the so-called “consistency profiles” in order to “characterize vague symbols,” and his view is a precursor of the idea of fuzzy set in the sense of Zadeh (1965). Fuzzy sets embody the notion of gradual predicates for which the idea of a precise boundary between situations where this predicate applies and situations where it does not is meaningless. However, Zadeh (1978) considered that vagueness covers both fuzziness and nonspecificity. He wrote:

Although the terms fuzzy and vague are frequently used interchangeably in the literature, there is, in fact, a significant difference between them. Specifically, a proposition, \( p \), is fuzzy if it contains words which are labels of fuzzy sets; and \( p \) is vague if it is both fuzzy and insufficiently specific for a particular purpose.

(Zadeh 1978, p. 396, note)

Zadeh’s introduction of fuzzy sets was not meant to be a contribution to the philosophy of vagueness. It was motivated by the need for a computational representation for linguistic terms appearing in statements, which are often intended to provide synthetic information about complex situations.

In this chapter, we also approach the discussion of vagueness from an information processing perspective, by focusing on knowledge representation aspects in the sense of Artificial Intelligence. The chapter systematically investigates information scenarios where forms of what could be called “vagueness” appear. We shall refer to the idea of a vague concept (or category) as soon as this concept partitions the universe of discourse (sometimes implicitly) into more than two regions. We investigate six basic scenarios that we identify as giving rise to situations of this kind. In Section 2, we first introduce the information framework common to all of the scenarios, and the corresponding notations. Sections 3–8 are devoted to the presentation and discussion of these
six scenarios. As emphasized in the conclusion, though, hybrid situations can be encountered, where several features of basic situations are found together.

2. The information framework

We use a general information-based framework, where objects are described by an agent in terms of attribute values and can be assigned to categories according to their properties. Such properties refer to subsets of attribute domains. Namely, let \( O \) be a finite set of objects or entities, and \( \mathcal{A} \) be a finite set of attributes applicable to these objects. The possible values of an attribute \( a \) in \( \mathcal{A} \) for the objects in \( O \) belong to the attribute domain \( D_a \). Therefore we shall understand each attribute \( a \) in \( \mathcal{A} \) as a mapping \( a:O \to D_a \). A property \( A \) regarding an attribute \( a \) will refer to the relationship between the objects and some classification of the attribute values in \( D_a \), as explained here.

For a two-valued (or equivalently Boolean) attribute \( a \), \( D_a \) contains two elements only, say \( \{y, n\} \). Thus, one can only speak of a property \( A \) and of its opposite \( \neg A \) with respect to attribute \( a \). The property \( A \) is true or not true for an object \( o \) in \( O \) depending on whether \( a(o) = y \) or \( a(o) = n \), respectively. Thus, each object satisfies either \( A \) or \( \neg A \). If we have complete information about all objects in \( O \) regarding attribute \( a \), then \( A \) is not perceived as a vague category. Let \( \text{Ext}(A) \) be the extension of \( A \) in \( O \), as perceived by the agent, i.e.,

\[
\text{Ext}(A) = \{ o \in O \mid a(o) = y \}; \quad \text{Ext}(\neg A) = \{ o \in O \mid a(o) = n \}.
\]

More generally, if the attribute domain contains more than two elements, the property \( A \) and its negation \( \neg A \) respectively refer to a pair of nonempty subsets \( Y_A \) and \( N_A \) of \( D_a \). For a classical property, \( Y_A \) and \( N_A \) form a partition of \( D_a \). The extensions \( \text{Ext}(A) = \{ o \in O \mid a(o) \in Y_A \} \), and \( \text{Ext}(\neg A) = \{ o \in O \mid a(o) \in N_A \} \), also form a partition of the set of objects. From now on, we shall identify properties \( A \) and \( \neg A \) with the subsets \( Y_A \) and \( N_A \) of \( D_a \), respectively, when no confusion is possible from the context.

It should be noted that, in daily practice, a property is associated with a label in a natural language. There is a context-dependent use of labels that is important for natural language understanding, but it will not be considered further in this chapter.

In the following sections, we study several variants of the above information framework: First, \( A \) and \( \neg A \) may not partition \( D_a \), because they are gradual properties. In the next variant, the agent may not know the extension of a property precisely, even if this extension exists. Another case is when the attribute domain \( D_a \) is equipped with a distance, and a notion of conceptual centrality can be introduced accordingly. In yet another setting, several agents may partially disagree on the extension of the property \( A \), resulting in global uncertainty. In the fifth scenario, some attribute values may not be well known for some objects. Finally, the attributes considered may not provide a sufficiently expressive language for characterizing some subsets of objects precisely.
The image contains a page from a document discussing classical vs. gradual properties in the context of attribute values. It explains that a property is classical if it creates an ordinary partition of the domain. It lists the Excluded Middle Law and Noncontradiction Law in the context of classical properties. It also mentions that these laws may fail in more general settings, leading to the concept of intuitionistic or paraconsistent properties. The text further elaborates on gradual properties and their natural language examples, such as 'young,' 'small,' and 'heavy,' which do not lead to a clear-cut binary partition. The text then defines gradual properties as those that are inherently gradual and discusses how they naturally induce a preordering on the set of objects. It provides examples, such as the classification of animals like 'bird' and 'penguin,' and discusses the use of numerical scales for attributes like height and age. The text also mentions the importance of considering the progressive failure of features when dealing with gradual properties and provides examples from the literature on fuzzy sets and cognitive categories.
just an ordered structure \((D_a, \geq_A)\), advocated by various authors [Finch (1981), Basu et al. (1992), Trillas and Alsina (1999), Lee, Yeung and Tsang (2002), Lee (2003)] is very elegant; however, it is very difficult to exploit for operational purposes when it comes to building a logic, given the lack of commensurateness between two such pre-orderings if they pertain to distinct properties.

3.2. Membership functions as total preorders

A richer representation scheme involves modeling the extensions of gradual properties by means of fuzzy sets [Zadeh (1965)]. In that case, to a structure \((D_a, \geq_A)\) we attach a membership function \(\mu_A : D_a \to [0, 1]\) preserving the ordering \(\geq_A\), that is, verifying \(\mu_A(u) \geq \mu_A(v)\) whenever \(u \geq_A v\), and mapping to 1 the maximal elements of \((D_a, \geq_A)\), or conversely, mapping the minimal elements to 0. Notice that when we do this, we are in fact enriching the knowledge representation setting. In effect, we are extending the possibly partial ordering \(\geq_A\) to a linear (and thus total) preorder. See Keefe (1998) for a negative view of the adequacy, for arbitrary vague concepts, of the assumption of such a linear extension, and its measurement on a continuous scale like \([0, 1]\). In the above example, the membership function for middle-aged enables every age to be compared with any other, simply by comparing their membership values. Moreover, we are also assuming that maximal elements of \((D_a, \geq_A)\) are fully compatible with \(A\) (and that the minimal elements are fully incompatible with \(A\)), thus providing landmarks for full membership (or complete lack of membership in the opposite case). These landmarks or anchor values cannot be expressed by means of a partial ordering alone. The preorder induced by a fuzzy set on the domain \(D_a\) partitions it into (possibly) infinitely many subsets \(\{u \in D_a \mid \mu_A(u) = \alpha\} \alpha \in [0, 1]\), in contrast with the binary partitions of classical properties. In fact, here we could replace \([0, 1]\) by any other linear, bounded, sufficiently discriminating scale. Observe that there is some relation between the nature of the attribute domain \(D_a\) and the possible number of levels in the membership scale. For modeling gradual properties, the membership scale needs to be (and naturally becomes) a continuum only if \(D_a\) is a continuum. Indeed, vagueness (in fact, fuzziness) naturally arises when one tries to represent a gradual property on a continuous referential \(D_a\). In particular, any classical partition-based representation of such a property leads to a sorites paradox. See Goguen (1969), Gaines (1977), and Copeland (1997) for fuzzy-set-based discussions of this paradox.

Focusing only on the boundaries of the membership scale, fuzzy sets naturally induce a tripartition of the attribute domain. Indeed, let \(Y_A = A^\circ\) be the core of \(A\), which comprises the elements that undisputedly belong to \(A\), i.e., \(A^\circ = \{u, \mu_A(u) = 1\}\). Similarly, \(N_A = (\neg A)^\circ = \{u, \mu_A(u) = 0\}\), assuming that fuzzy set complementation agrees with classical complementation for the extreme values in the scale. Then, for a genuine fuzzy set, we have the strict inclusion \(Y_A \cup N_A \subset D_a\). Thus, we can define a set of borderline elements as \(B_A = D_a - (Y_A \cup N_A)\). Also, note that the supports of \(A\) and \(\neg A\), namely \(A_s = \{u, \mu_A(u) > 0\}\) and \((\neg A)_s = \{u, \mu_A(u) < 1\}\), are not disjoint. Fuzzy sets violate the excluded middle and noncontradiction laws (EM) and (NC): on the one
hand, the support of a fuzzy set and the support of its complement overlap, while the union of their cores does not cover the referential domain. This violation emphasizes the fact that, with genuine fuzzy sets, a clear-cut boundary between $A$ and its complement does not exist. This view of vagueness is quite similar to that introduced by Black (1937), and is close to Alston’s (1964) definition of degree of vagueness. The pair $(Y_A, Y_A \cup B_A)$, was called an ensemble flou by Gentilhomme (1968), who viewed $Y_A$ as the set of central elements of $A$ and $B_A$ as the set of peripheral elements. See Lakoff (1987) and Smithson (1987) for further discussion of gradations in categories.

3.3. Fuzzy sets and similarity to prototypes

When the agent is able to measure how close or similar one element of the domain $D_a$ is to another with respect to the attribute $a$, one can propose the following computation of membership degrees [Ruspini (1991)]. Assume the agent is provided with a closeness relation $S: D_a \times D_a \to [0, 1]$, verifying at least $S(u, u) = 1$ for all $u \in D_a$, where $S(u, v) = 1$ means that $u$ and $v$ are indistinguishable, $S(u, v) = 0$ means that $u$ and $v$ have nothing in common, and $S(u, v) > S(u, v')$ means that $u$ is more similar to $v$ than to $v'$. Then, given $A^\circ$ and $(\neg A)^\circ$, which can be seen as (proto)typical values defining $A$ and $\neg A$, respectively, one can define the degree $\mu_A(u)$ in which a value $u$ belongs to $A$, as the extent to which $u$ is close or similar to some typical value of $A^\circ$. In some sense, we are identifying $A$ with those values that are close to (or within) $A^\circ$. We proceed similarly for $\mu_{\neg A}(u)$. According to this view, we can define for all $u$ in $D_a$:

$$
\mu_A(u) = \sup\{S(u, v) \mid v \in A^\circ\}; \quad \mu_{\neg A}(u) = \sup\{S(u, v) \mid v \in (\neg A)^\circ\}.
$$

Note that $\mu_A(u) = 1$ for all $u \in A^\circ$ and $\mu_{\neg A}(u) = 1$ for all $u \in (\neg A)^\circ$, but in principle nothing prevents $\mu_A(u) > 0$ for some $u \in (\neg A)^\circ$, or $\mu_{\neg A}(u) > 0$ for some $u \in A^\circ$ from existing. See Osherson and Smith (1981) for a critical discussion of a fuzzy-set-based approach to prototype theory, and Zadeh’s (1982) reply. This view may be related to Weston’s (1987) idea of approximate truth as reflecting a distance between a statement and the ideal truth. It is also related to Niiniluoto’s (1987) notion of truth-likeness and to similarity-based reasoning, as developed by Dubois et al. (1997).

3.4. Set-theoretic operations

Fuzzy set theory has developed an algebraic framework for defining truth-functional set-theoretic operations extending classical set operations. Intersections and unions are then defined point-wise\(^1\). Depending on the operations used for defining the fuzzy set

\(^1\) That is, $\mu_{\cap A}(u) = \mu_A(u) \ast \mu_B(u)$ and $\mu_{\cup A}(u) = \mu_A(u) \perp \mu_B(u)$, where the two-place operations $\ast$ and $\perp$ on $[0, 1]$ are commutative, associative, monotonically nondecreasing, with appropriate boundary conditions (for intersection, $1 \ast x = x$; for union, $0 \perp x = x$). They are named triangular norms and co-norms, respectively [e.g., Alsina, Trillas and Valverde (1983)]. Moreover, $\mu_{\neg A}(u) = 1 - \mu_A(u)$ for complementation.
intersection, equalities (EM) and (NC) may fail or hold. However, the excluded middle and noncontradiction laws, and idempotence \((A \cap A = A = A \cup A)\) cannot be satisfied at the same time: idempotence holds only with min- and max-based intersection and union, while (EM) and (NC) are preserved for non-idempotent operations such as Lukasiewicz connectives\(^2\) (although the supports of \(A\) and \(\neg A\) overlap!). This necessarily leads to structures weaker than Boolean algebras for fuzzy sets.

Clearly, truth-functionality of connectives is a nice property for computation to have, when possible. This simplicity can be obtained when we assume a unique membership scale for all concepts on \(D_a\). However, it is not easy – even if not impossible – to define (binary) connectives for concepts described with only two preorders \((D_{a'}, \geq_{a'})\) and \((D_{a''}, \geq_{a''})\) on the same domain. See Lee (2003) for an investigation of intersection and union connectives in this setting from the perspective of decomposability and ordinal conjoint structure in measurement theory. In any case, the most natural definitions are:

\[
\begin{align*}
\forall u \ni \neg A \cap B v & \iff \forall u \ni A v \land \forall u \ni B v \quad \text{(Pareto-ordering)} \\
\forall u \ni \neg A \cup B v & \iff \forall u \ni A v \lor \forall u \ni B v.
\end{align*}
\]

However the intersection will be very poorly discriminant, since for only a few pairs \((u, v)\) will it be true that \(\forall u \ni \neg A \cap B v\). Besides, \(\forall u \ni \neg A \cup B v\) is generally not transitive (its strict part will be neither transitive nor acyclic) and its transitive closure may very well be trivial. This is clearly related to Arrow’s (1963) impossibility theorem in the social sciences.

The truth-functionality assumption of membership functions is not without controversy. For instance, the similarity-based model presented above will be truth-functional for disjunction only. Moreover, this assumption may be found to be too simplistic for an accurate account of vagueness-originated phenomena; see Sanford (1975) for a critical discussion.

In the discussion above, we have referred to fuzzy set operations on a single attribute domain. Although properties, whose definition involves several attribute domains make the discussion of vagueness more complicated, they do not contribute any important new features to the discussion of vagueness in relation to the idea of gradual properties.

### 3.5. Graduality is a useful form of vagueness

When vagueness results from the use of gradual properties, it should not be seen as a defect to be remedied, but rather as a desirable capability by the language to capture the idea of typicality, and to interface linguistic categories with a continuum of attribute values (usually numerical), without introducing arbitrary discontinuities. This capability is accounted for by fuzzy set representation. Remember that the other vagueness scenarios, presented in the following sections, assume classical (nongradual) properties.

\[^2\] \(x \cdot y = \max(0, x + y - 1)\) and \(x \perp y = \min(1, x + y)\).
Some philosophers and logicians, e.g., Haack (1996), Parikh (1983), and Tyle [cited by Copeland (1997)] have pointed out the problem of “inappropriate precision” inherent in fuzzy set membership functions, which they find paradoxical when dealing with vagueness. However, the scenario considered in this section deals with the modeling of graduality, or partiality, which is mainly based on the idea of ordering, and which has nothing to do with imprecision. The notion of partial truth, as put forward by Lukasiewicz (1930), leads to a change in the very notion of proposition. The definition of proposition is a matter of convention, as stressed by De Finetti (1936):

Propositions are assigned two values, true or false, and no other, not because there “exists” an a priori truth called “excluded middle law,” but because we call “propositions” logical entities built in such a way that only a yes/no answer is possible.

Fuzzy sets deal with many-valuedness in a logical format; they are not primarily concerned with uncertainty or belief. Contrary to what the terminology (“vague,” “fuzzy”) may suggest, gradual predicates allow for a refined model of categories, more expressive than the Boolean setting, and reflecting the common usage of some words in the form of underlying preferred meanings or default typicality orderings in the situations they refer to. Membership functions are just convenient context-dependent numerical representations of this ordering. Gradual propositions contain more information than all-or-nothing ones. But the problem of the measurement of membership functions makes sense, and is discussed in the fuzzy set literature [see, e.g., Turksen and Bilgic (2000), Marchant (2004)].

4. Precisely defined vs. poorly defined properties

An agent may not be able to precisely delimit the extension of a clear-cut property $A$. By an imprecisely delimited extension, we mean the existence of a borderline region in $D_a$, where there exist elements for which the agent cannot say whether they should be classified as $A$ or $\neg A$. This is also called semantic ambiguity. Here, vagueness results from a lack of knowledge of the precise extension of property $A$, rather than from the lack of complete information regarding some attribute values of objects (for the latter scenario, see Section 7). Thus the most elementary representation of this situation, for properties that are not a matter of degree, is the partition of $D_a$ into the three subsets $Y_A$, $N_A$, and $B_A = D_a - (Y_A \cup N_A)$

where $Y_A$ (or conversely, $N_A$) is the set of attribute values that the agent can classify as belonging to $A$ (or conversely, $\neg A$) without any hesitation. $B_A$ is the borderline (uncertainty) area containing the real boundary of $A$. It is the set of attribute values that the agent can assign neither to $A$, nor to $\neg A$.

Some scholars, denying the existence of intrinsic graduality, model predicates like ‘young’ in this way. Indeed, the set $Y_A$ of elements with sure membership sounds like
the set $A^\circ$ of prototypical elements of a fuzzy set. However, in this section, we are considering the hypothetical situation of a classical property, whose precise meaning (i.e., the extension of the property) is not known by the agent, who is unsure whether some value or element satisfies the property.

The idea of subdefinite sets suggested by Narin’yani (1980) also acknowledges the fact that an agent may have only partial knowledge of the extensions of $A$ and $\neg A$. In that case, the nonmembership of an element in one set does not determine its membership in the complement. Thus, a subdefinite set $S$ is a pair $(A^+, A^-)$ of disjoint subsets of elements that definitely belong or definitely do not belong to $S$, together with some piece(s) of information on the cardinality of these subsets.

### 4.1. Classification ambiguity

When asked whether a certain value $u$ in the domain $D_a$ satisfies the property $A$ or not, an agent may express his belief in the membership or nonmembership in $A$ of values in $D_a$, by means of an uncertainty measure $g_A^u: 2^{\{y, n\}} \rightarrow [0, 1]$ for each $u$ in $D_a$, where $y$ and $n$ stand for “belonging to $A$” and “not belonging to $A$,” respectively. In such a case, $g_A^u$ induces two fuzzy sets on $D_a$ with membership functions defined by $\mu_A(u) = g_A^u(y)$ and $\mu_{\neg A}(u) = g_A^u(n)$. A reasonable condition is that $g_A^u(n) \leq g_A^u(y)$, and vice versa. The sets $Y_A$ and $N_A$ correspond to cases when $g_A^u(y) = 1$ and $g_A^u(n) = 1$, respectively. Other elements, with uncertain membership, belong to the boundary $B_A$ of $A$. Obviously, if the measures $g_A^u$ are probabilities verifying the above condition, then $\mu_A(u) = 1 - \mu_{\neg A}(u)$, and $\mu_A$ is similar to a likelihood function $P(A|u)$. The interpretation of membership functions as conditional probabilities $P(A|u)$ was stressed by Cheeseman (1986) and Hisdal (1988). Coletti and Scozzafava (2004) show that membership functions can be then cast in the theory of coherent conditional probability, which goes back to De Finetti. The work of Giles (1988) can be viewed as addressing the same subjective probability trend, whereby a membership grade is interpreted in terms of betting rates pertaining to (Boolean) membership.

### 4.2. Vagueness as limited perception

Parikh’s (1983) view can be related to the above representations. In his opinion, the idea of vagueness stems from a perception problem, namely the difficulty of defining (crisp) predicates on “observationally connected spaces” (e.g., colors) with insufficiently separated elements. Thus, rather than advocating a fuzzy set modeling in such a case [e.g., Kay and McDaniel (1975)], Parikh considers that the difficulty in assigning borderline elements or values to $A$ (or to $\neg A$) is due to an inability to discern or distinguish between them, since they are too close. So the boundary between the extensions of $A$ and $\neg A$ is poorly known, even if there are elements that can be considered as clearly belonging to $A$. More formally, suppose two elements $u$ and $v$ in $D_a$ are indiscernible as soon as $d(u, v) \leq e$, where $d$ is a distance function on $D_a$ and $e$ is an indiscernibility threshold. Then, each element $u$ in $D_a$ is perceived as the subset $\{u\} = \{v, d(u, v) \leq e\}$. So
$Y_A = \{ u, [u] \subseteq A \}$, and $N_A = \{ u, [u] \subseteq \neg A \}$. Any elements $u$ in $A$ and $v$ in $\neg A$ such that $d(u, v) \leq e$ will be perceived as lying in the borderline area $B_A$.

Another case of the same kind, where a borderline area may occur, is when the attribute range $D_a$ is replaced by clusters forming a partition of similar elements. This is the case when one considers a coarsening (or granulation) of the attribute range (e.g., measuring heights in centimeters instead of millimeters). There is a classical equivalence relation on $D_a$ and each element $u$ in $D_a$ is perceived as the equivalence class $[u]$. The partition $(A, \neg A)$ of $D_a$ is again perceived as a trichotomy, as before, here due to the use of a coarse scale.

4.3. Supervaluations

In all of the above settings, although the agent is not able to locate the boundary between $A$ and $\neg A$, he still assumes that the excluded middle and noncontradiction laws (EM) and the excluded-middle and contradiction laws, (EM) and (NC), hold. Indeed, Fine (1975), advocating the idea of “super-truth,” proposes that statements about a vague predicate are “superture” if and only if they hold for all possible ways of making the predicate precise [see also van Fraassen (1969), Keefe (2000)]. This assumption enables all classical logical relationships between a vague predicate $A$ and its negation $\neg A$ to be preserved. See Sanford (1976, 1979) for various points of view about the idea of super-truth. This view appears to be similar to Williamson’s (1994) view of vagueness. For Williamson, $Y_A$ corresponds to those elements that are “clearly $A$.”

4.4. Ill-known partial membership

Semantic ambiguity may also result from gradual properties. The imprecision of $\mu_A$ can be captured by a type-2 fuzzy set (Mizumoto and Tanaka (1976), where $\mu_A(u)$ is itself a fuzzy set of $[0, 1]$. A particular case called an “interval-valued fuzzy set” occurs when $\mu_A(u)$ is an ordinary subinterval of $[0, 1]$ [Grattan-Guiness (1975)]; such cases are also called vague sets by Gau and Buehrer (1993). Atanassov (1986, 1999) extends Narin’yi subdefinite sets by defining a so-called “intuitionistic fuzzy set (IFS)” $iA$ as a pair of membership functions $(\mu_{A+}, \mu_{A-})$, where $\mu_{A+}(u)$ is the degree of membership of $u$ in $iA$ and $\mu_{A-}(u)$ is its degree of nonmembership. The two membership functions are supposed to verify the constraint $\mu_{A+}(u) + \mu_{A-}(u) = 1$. The name “intuitionistic” stems from this inequality, which is supposed to express a rejection of the excluded middle law, but since negation is involutive in this theory (it amounts to swapping $\mu_{A-}$ and $\mu_{A+}$), the name is misleading. In fact Atanassov’s construct is isomorphic to interval-valued fuzzy sets. See Bustince and Burillo (1996) and Deschrijver and Kerre (2003) for example.

5. Refining precisely defined properties using closeness relations

The situation considered in the previous section can be viewed as a case, where the information that would enable one to decide between $A$ and $\neg A$ is poor or incomplete.
With rich information, precisely delimited extensions may also lead to a trichotomy of \( D_a \) if it is possible to measure how close any two attribute values are to each other. Here, we do not assume any perception deficiency: the agent can always distinguish between any two distinct attribute values \( u \) and \( v \), no matter how close, so that \( A \) and \( \neg A \) are well known and form a partition of \( D_a \). Consider there is a graded closeness or similarity relation \( S: D_a \times D_a \rightarrow [0, 1] \), which is

- reflexive: \( S(u, u) = 1 \),
- symmetric: \( S(u, v) = S(v, u) \),
- separating: \( S(u, v) < 1 \) whenever \( u \neq v \).

\( S(u, v) \) is all the greater as \( u \) and \( v \) are close to each other. It is a monotonically decreasing function of a distance. The separating property is essential here to indicate that the agent perceives the boundary of \( A \) perfectly. We can then define the fuzzy set of central elements of \( A \) and \( \neg A \) by means of the membership functions

\[
\mu_{YA}(u) = 1 - \sup\{S(u, v) \mid v \in \neg A\}, \quad \mu_{NA}(u) = 1 - \sup\{S(u, v) \mid v \in A\}.
\]

Here, the Boolean representation is refined by making \( Y_A \) and \( N_A \) gradual. Elements not in \( A \), but outside the core of \( N_A \), lie in the vicinity of \( A \), and can be used for interpolation reasoning [Ruspini (1991)]. If \( \mu_{Y_A}(u) > 0 \), then necessarily \( u \in A \), so \( Y_A \) is indeed included in \( A \). Moreover, \( u \) is a fully central element for \( A \) (i.e., \( \mu_{Y_A}(u) = 1 \)) as soon as \( u \) is totally dissimilar from some element \( v \) of \( \neg A \) (i.e., \( S(u, v) = 0 \)).

In some sense, this situation is opposite to the one described in the previous section. There, assuming \( A \) is a binary property, we could explain the lack of knowledge about its boundary using an indiscernibility relation induced by a perception threshold, or some uncertainty measure. Making this indiscernibility gradual, we formally get the same expressions as above for computing valued (fuzzy) counterparts to \( Y_A \) and \( N_A \), but the meaning is very different. In the previous section, the boundary region contained elements of uncertain membership. Here, on the contrary, \( A \) is well defined but we are interested in describing central elements of \( A \), which lie far away from elements of \( \neg A \). The similarity relation enables a membership function for the fuzzy set of central elements of \( A \) to be derived.

For instance, consider exam grades in the range \([0, 20]\). It is perfectly well known that students with marks of 10 or higher pass, while the rest fail. Yet the really successful students are those, whose marks are much higher than 10, while the really unsuccessful ones are those whose marks are much lower than 10. Here the graduality of \( Y_A \) and \( N_A \) makes the representation more expressive, and does not convey any idea of uncertainty.

### 6. Single agent vs. multiple agents

Vagueness also results when different extensions of a property \( A \) (and \( \neg A \)) are provided by a set of agents, even if each agent perceives \( A \) as a classical property. Indeed, let \( Y_{Ai} \) and \( N_{Ai} = \neg Y_{Ai} \) be the dichotomic (agent-dependent) representations of \( A \) for agent
Assume for the sake of simplicity that they are classical extensions. This situation implicitly generates a partition of $D_a$ into three regions:

$$Y_A = \bigcap_i Y_{Ai}, N_A = \bigcap_i N_{Ai}, B_A = D_a - \left( \bigcap_i Y_{Ai} \cup \bigcap_i N_{Ai} \right).$$

In extreme cases, we may have $\bigcap_i Y_{Ai} = \emptyset$ or $\bigcap_i N_{Ai} = \emptyset$ if the agents are completely inconsistent.

In the case of multiple agents, it is natural to try to summarize the different points of view. One way to do this is to attach to each $(Y_{Ai}, N_{Ai})$ the weight $m(Y_{Ai})$ given by the proportion of agents, who consider that the correct extension of $A$ is $Y_{Ai}$. Then $\sum_i m(Y_{Ai}) = 1$.

On the basis of the proportion of individuals, who think that $A_i$ properly expresses $A$, we can define the grade of membership of $u$ to the (agent-dependent) concept $A$. $m_{A}(u)$ estimates the extent to which the value $u$ is globally compatible with the meaning of $A$. This is formally expressed in the form of a random set or equivalent to a body of evidence in Shafer’s sense. It corresponds exactly to Zadeh’s definition of a fuzzy set, as soon as the family $\{Y_{Ai}, m(Y_{Ai})\}$ is a nested family so that the knowledge of the membership function $\mu_{A}$ is equivalent to that of the probabilities $m(Y_{Ai})$ [see Dubois and Prade (1989, 1990)]. Of course, this nested property is seldom observed in practice, since the $Y_{Ai}$s come from different agents. However consonant (nested) approximations of dissonant bodies of evidence exist [Dubois and Prade (1989)]. These are especially good when $\bigcap_i Y_{Ai} \neq \emptyset$, a usually satisfied consistency requirement, which expresses that there exists at least one value in $D_a$ that is totally compatible with the concept for everybody in a given context. Hence a fuzzy set, with membership function $\mu_A: D_a \rightarrow [0, 1]$ can always be used as an approximation of a random set. Such a construct can be used for measuring the membership function $\mu_{A}$ of a fuzzy set $A$ (e.g., ‘young’) in a given context. Then $A$ is a fuzzy set and $Y_{Ai}$ is a crisp realization of the idea of fuzzy set $A$ for an individual $i$.

A simpler but related experiment consists of asking each agent $i$, for each value $u$, if $u$ is or is not in the extension of $A$. Then $\mu_{A}(u)$ would just reflect the proportion of individuals, who answer that $u$ is in the extension of $A$; psychologists have used this to obtain fuzzy set membership functions [e.g., Hersh and Caramazza (1974)]. Then $\mu_{A}$ is obtained via a likelihood function $P(A|u)$. This view is also a translation of Cheeseman’s (1986) definition of vagueness.

These two probability-oriented views (random sets and likelihood functions) of fuzzy sets are not antagonistic and can be reconciled. The random set view corresponds to an experiment whereby individuals are asked to point out a single crisp subset $Y_{Ai} \subseteq U$ that best represents some fuzzy concept $A$. The weight $m(Y_{Ai})$ represents the proportion of individuals for which $A$ is best described by $Y_{Ai}$. It makes sense to relate this experiment to the likelihood Yes-No experiment provided that, if an individual chooses $A_i$ as representing $A$, it means that, in the other experiment, he or she would answer yes
to the question “is \( u \) an ‘A’?" if and only if \( u \in A_i \). Then, as pointed out by Dubois and Prade (1990), the likelihood function and the random set view are in agreement, i.e.,

\[
P('A'|u) = \sum_{u \in Y_A} m(Y_A).
\]

When one performs logical operations on the representations of \( A \) and \( B \), there are two possibilities: (i) perform the operation for each agent \( i \) on \( Y_{Ai} \) and \( Y_{Bi} \) and then compute the resulting membership functions, or (ii) perform the operation on the membership functions (the summarized views) of \( A \) and \( B \). Clearly, the first option is most respectful of the agents.

Another approach defines multiple-agent vagueness directly on the set of objects \( O \) without making the attribute scale explicit [Lawry (2004)]. More precisely, it starts with a term set \( t \) of predicate symbols \( \{A_k, k = 1, \ldots, n\} \), which describe linguistic values of some attribute. For a given object \( o \in O \), each agent points out a subset \( S_i(o) \) of terms, each of which is considered compatible with the object. Let \( m(S_i) \) be the proportion of agents, who consider that the set \( S_i(o) \) of terms is compatible with object \( o \). The fuzzy extension of a predicate \( A \) on the set of objects is defined as

\[
\mu_{\text{Ext}(A)}(o) = \sum_{S_i \in S} m(S_i(o)).
\]

The advantage of this approach is that it is possible to apply it to abstract predicates whose underlying measurement scales are not obvious.

The idea of agent-dependent concepts applied to a gradual property gives rise to the notion of probabilistic set [Hirota (1981), Czogala and Hirota (1986)], where membership degrees are known only through probability distributions. Finally, Halpern (2004) envisages an approach to vagueness combining both the idea of variability of the meaning across several agents and limited perception by each agent, modeled by nontransitive indiscernibility relations.

7. Ill-known attribute values and twofold sets

In most of the previous situations, vagueness stemmed from peculiarities of the way an agent perceived an attribute scale. In each case, the vagueness of the extension of \( A \) on the set of objects was a direct reflection of the vagueness of the representation of the property \( A \) on the attribute scale. Here, we assume \( A \) is perceived as classical, so that \( Y_A \) and \( N_A \) partition the attribute domain \( D_a \) in the classical sense. However, the agent’s knowledge of the values of attribute \( a \) for objects is uncertain, or simply incomplete. The set of objects, which are \( A \) is then ill known. For instance, if for each object \( o \), \( p_a(o) \) denotes the subjective probability distribution of the possible values for \( a(o) \) according to the agent, we note that

\[
\text{Prob}(a(o) \in A) = \sum_{u \in Y_A} p_a(o)(u);
\]
and then define the fuzzy set of objects satisfying \( A \) as having the membership function

\[
\mu_{\text{Ext}(A)}(o) = \text{Prob}(a(o) \in A).
\]

Clearly, the boundary between objects that satisfy \( A \) and those that do not is blurred. Only for well-known objects do we know that \( o \in \text{Ext}(A) \) or not. The excluded middle and noncontradiction laws on \( D_a \) imply

\[
\mu_{\text{Ext}(A)}(o) + \mu_{\text{Ext}(\neg A)}(o) = 1.
\]

When the information is poorer, we can still define the extensions. Suppose all the agent knows about \( o \) is that \( a(o) \in I(o) \), a subset of \( D_a \). Then the agent knows for sure that “\( o \) is \( A \)” if and only if \( I(o) \subseteq Y_A \) and that “\( o \) is not \( A \)” if and only if \( I(o) \subseteq N_A \). If neither condition holds, \( o \) is borderline: it is neither in \( \text{Ext}(A) \) nor in \( \text{Ext}(\neg A) \), which are disjoint and no longer form a partition of \( O \).

More generally, for each object there is a possibility distribution \( \pi_{a(o)} \) describing the more or less possible values of \( a(o) \) for each object. Then one may compute [Dubois and Prade (1987)]:

- to what extent it is possible that object \( o \) has property \( A \):

\[
\forall o \in O, \quad \Pi(a(o) \in A) = \sup_{u \in A} \pi_{a(o)}(u) = 1 - \mu_{\text{Ext}(\neg A)}(o)
\]

- to what extent it is certain that object \( o \) has property \( A \):

\[
\forall o \in O, \quad N(a(o) \in A) = 1 - \Pi(a(o) \in \neg A) = \mu_{\text{Ext}(A)}(o)
\]

Since \( N(a(o) \in A) > 0 \) implies \( \Pi(a(o) \in A) = 1 \), the (fuzzy) set of objects \( \text{Ext}(A) \), which are more or less certainly \( A \), is disjoint from the (fuzzy) set of objects \( \text{Ext}(\neg A) \), which are more or less certainly \( \neg A \). In this situation, the set of objects that have property \( A \) is poorly known because of the imprecise descriptions of these objects. This gives rise to a trichotomistic structure, not in the attribute domain \( D_a \) but in the set of objects \( O \): the set of objects \( A^* \), which are certainly \( A \) (such that \( \mu_{\text{Ext}(A)}(o) = 1 \)), the set \( (\neg A)^* \) of objects, which are certainly \( \neg A \) (such that \( \mu_{\text{Ext}(\neg A)}(o) = 1 \)), and the boundary set \( O - \{A^* \cup (\neg A)^*\} \).

8. Approximately described sets

Yet another situation, where vagueness appears in the set of objects \( O \), rather than in the attribute scale, is when the language induced by the set of attributes \( A \) does not allow one to accurately describe some subsets \( E \) of \( O \). This occurs when several objects share the same description. In such a case, we can only define a lower and an upper approximation of a given set \( E \) of objects, since objects with the same attribute values for an attribute \( a \) are indiscernible from the point of view of this attribute. This is actually the starting point of rough set theory [Pawlak (1991)].

Consider first an attribute \( a \) and define, for each \( u \in D_a \), the equivalence classes

\[
[O]_u = a^{-1}(u) = \{ o \in O | a(o) = u \}.
\]
Note that for \( u \neq v \), \([O]_u \cap [O]_v = \emptyset\). This can be generalized to a subset \( \mathcal{B} = \{a_1, a_2, \ldots, a_r\} \) of attributes, taking \([O]_u = \{ o \in O \mid \mathcal{B}(o) = u \}\) to define an equivalence class for each \( r \)-tuple \( u = (u_1, \ldots, u_r) \in D_{a_1} \times \cdots \times D_{a_r} \) where \( \mathcal{B}(o) = u \) stands for \( a_1(o) = u_1, \ldots, a_r(o) = u_r \). We can then define the lower and upper approximations, respectively, of \( E \) in \( O \) as:

\[
E_* = \{ o \in O \mid [O]_{\mathcal{B}(o)} \subseteq E \}, \quad E^* = \{ o \in O \mid [O]_{\mathcal{B}(o)} \cap E \neq \emptyset \}.
\]

Then clearly, \( E_* \subseteq E \subseteq E^* \). This again leads to a trichotomic structure: \((E_*, E^* - E_*, \neg(E^*))\).

The rough set model can be enriched by dealing with fuzzy indiscernibility relations or fuzzy partitions instead of equivalence classes, or by approximating fuzzy subsets in \( O \) (rather than classical subsets only). See Dubois and Prade (1992) for a combination of the rough set scenario and gradual predicates. Boixader, Jacas and Recasens (2000) discuss the approximation of fuzzy sets using similarity relations.

9. Concluding remarks

Overall, there are basically three types of approaches to vagueness (see Table 1):

- Approaches that admit from the start that propositions or predicates may fail to be Boolean. These correspond essentially to the first scenario underlying the fuzzy set paradigm and many-valued logics. The intended meaning of propositions is richer than the use of Boolean variables to represent them might suggest. This situation is precisely what Zadeh calls fuzzy (and not vague). Scenario 3 (in Table 1) can be viewed as belonging to the same school of thought, since properties are defined as essentially Boolean, and then some refinement separating central from peripheral elements is introduced by means of the underlying distance on the attribute scale. Again,

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Overview of vagueness scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scenarios</strong></td>
<td><strong>Boolean notion</strong></td>
</tr>
<tr>
<td>1. Gradual predicates</td>
<td>No</td>
</tr>
<tr>
<td>2. Uncertain boundaries</td>
<td>Yes</td>
</tr>
<tr>
<td>3. Closeness</td>
<td>Yes, but refined</td>
</tr>
<tr>
<td>4. Multiagent</td>
<td>Yes</td>
</tr>
<tr>
<td>5. Ill-known objects</td>
<td>Yes</td>
</tr>
<tr>
<td>6. Ill-described objects</td>
<td>Yes</td>
</tr>
</tbody>
</table>
it is an enrichment of the Boolean representation. All other approaches preserve the Boolean representation convention, and consider vagueness (and graduality) as stemming from deficient information.

- Approaches that claim that the boundary between values satisfying a proposition exist but is ill-known due to the limited perception of one agent or the conflicting views of several agents. These approaches correspond to scenarios 2 and 4, respectively. Ill-defined properties induce a vague classification of objects. Moreover, since properties are still Boolean even if their boundaries are not well-known, the basic laws of classical logic are retained, such as the laws of excluded middle and noncontradiction. This is a natural view of vagueness understood as semantic ambiguity.

- Approaches, where the information defect lies in the difficulty of describing objects by means of suitable attributes, either because the attribute values of objects are ill-known (as in scenario 5) or because there are not enough attributes to ensure a bijection between the set of objects and the set of descriptions (scenario 6). In that case, vagueness is only reflected in a limited capacity to classify objects, but it does not affect the representation of properties on attribute ranges.

Clearly, these basic scenarios can be combined into more complex ones, where several key vagueness-generating features are present at the same time. A natural follow-up of this investigation is the study of set-theoretic connectives that may be used to logically combine vague properties in light of the above scenarios. This issue is the major cause of controversies in the philosophical and computer science literature [see Elkan (1994)]. We believe that some of these controversies stem from misunderstandings due to a failure to acknowledge the existence of several types of vagueness phenomena and the temptation to comment on proposals made within one scenario in the context of another scenario. A typical confusion is between degrees of truth (underlying scenario 1) and degrees of uncertainty (underlying scenarios 2 and 4).

References


This page intentionally left blank
Chapter 41

A SMOOTH INTRODUCTION TO SYMBOLIC METHODS FOR KNOWLEDGE DISCOVERY

AMEDEO NAPOLI

Loria, France

Contents

Abstract \hspace{1cm} 914
1. Introduction \hspace{1cm} 915
2. Methods for KDD \hspace{1cm} 916
   2.1. An introductory example \hspace{1cm} 916
   2.2. Data mining methods \hspace{1cm} 917
3. Lattice-based classification \hspace{1cm} 918
4. Frequent itemset search and association rule extraction \hspace{1cm} 920
   4.1. Frequent itemset search \hspace{1cm} 921
   4.2. Association rule extraction \hspace{1cm} 923
5. Applications \hspace{1cm} 923
   5.1. Mining chemical reaction database \hspace{1cm} 924
      5.1.1. The chemical context \hspace{1cm} 924
      5.1.2. Mining of a reaction database \hspace{1cm} 925
      5.1.3. Discussion \hspace{1cm} 926
   5.2. An experiment in biology \hspace{1cm} 927
   5.3. An introduction to Web mining \hspace{1cm} 928
6. Discussion \hspace{1cm} 930
7. Conclusion \hspace{1cm} 930
References \hspace{1cm} 931
Abstract

In this chapter, we present a smooth introduction to symbolic methods for knowledge discovery in databases (KDD). The KDD process extracts from large databases information units that can be interpreted as knowledge units to be reused. This process has three major steps: the selection and preparation of data, the data mining operation, and finally the interpretation of the extracted units. The process may take advantage of domain knowledge embedded in domain ontologies, which may be used at every step of the KDD process. We describe three symbolic methods for KDD: lattice-based classification, frequent itemset search, and association rule extraction. Then, we present three applications of the KDD process, and end the chapter with a discussion of the main characteristics of KDD.
1. Introduction

Knowledge discovery in databases (KDD) can be likened to the process of panning for gold in rivers: the gold nuggets that are sought are knowledge units, and the rivers are the databases under study. Huge volumes of data – and particularly documents – are available, without any specified intended use. A fundamental question is whether there is anything interesting in these data and, if so, how we can find methods to extract these “interesting things.” The KDD process consists in processing a huge volume of data in order to extract knowledge units that are non–trivial, potentially useful, significant, and reusable. Generally, the KDD process is iterative and interactive, and controlled by an expert in the data domain, called the analyst, who is in charge of guiding the extraction process, on the basis of his or her objectives and domain knowledge. The analyst selects and interprets a subset of the units to build “models” that will be further considered as knowledge units with a certain plausibility. The KDD process is based on three major steps: (i) the data sources are prepared to be processed, (ii) they are mined, and (iii) finally, the extracted information units are interpreted so they can become knowledge units. These units are in turn embedded within a representation formalism to be used within a knowledge-based system. The KDD process may also be understood as a process that turns data into information and then into knowledge (see Figure 1), considering the following equations [Schreiber et al. (1999); Wille (2002)]: (i) data = signs + syntax, (ii) information = data + meaning, (iii) knowledge = information (syntax and semantics) + ability to use information.

The KDD process is performed within a KDD system that is composed of the following elements: the databases, the symbolic or numerical data mining modules, and the interfaces for interactions with the system, e.g., editing and visualization. Moreover, the KDD system may take advantage of domain knowledge embedded within an ontology relative to the data domain. Closing the loop, the knowledge units extracted by the KDD system must be represented in an adequate representation formalism, and then they may be integrated into the ontology, to be reused for problem-solving needs in application domains such as agronomy, biology, chemistry, medicine, etc.

This chapter constitutes a smooth introduction to KDD that will focus on the so-called symbolic methods in knowledge discovery. There are a number of general books that can productively be used to understand KDD principles and the use of KDD methods, both historical research books such as Fayyad et al. (1996) and Michalski, Bratko and Kubat (1998) and more recent textbooks such as Dunham (2003), Han and Kamber (2001), Hand, Mannila and Smyth (2001), and Witten and Franck (2000), which is associated with the Weka system. In this chapter, we present three symbolic methods for KDD, namely lattice-based classification,
frequent itemset search, and association rule extraction. Then, we detail some applications of the KDD process. The chapter ends with a discussion and conclusion.

2. Methods for KDD

2.1. An introductory example

First, let us examine what can be expected of the application of data mining methods to data. Let us consider a Boolean matrix \( M_{ij} \), also called a formal context, where the rows stand for customers, and the columns for products bought by the customers (see Figure 2): \( M_{ij} = 1 \) whenever the \( i \) buys product \( j \). In real-world formal contexts, such a Boolean matrix may have thousands of columns, and millions of lines. From this formal context, one may extract the following units:

- The set \( X = \{ \text{beer, sausage, mustard} \} \) has a frequency of \( \phi(X) = 0.4 \), i.e., four individuals out of 10 buy the three products together. In the same way, the set \( X' = \{ \text{beer, sausage} \} \) has a frequency of \( \phi(X') = 0.6 \). The set \( X \) (or \( X' \)) may be interpreted as indicating that 40% (or 60%) of the customers buy the products in \( X \) (or in \( X' \)) at the same time.
- Moreover, the rule \( R = \{ \text{beer and sausage} \rightarrow \text{mustard} \} \) may be extracted from the sets \( X \) and \( X' \) (i.e., \( X' \rightarrow X \setminus X' \) where \( X \setminus X' \) denotes the set \( X \) without \( X' \)), with a
Fig. 2. An example of a Boolean matrix representing transactions between customers (C) and products (P).

<table>
<thead>
<tr>
<th>Customers/Products</th>
<th>chips</th>
<th>mustard</th>
<th>sausage</th>
<th>soft drink</th>
<th>beer</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C10</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

confidence level of 0.66, i.e., if a customer buys sausage and beer, then the probability that he will also buy mustard is 0.66 (out of six customers who buy sausage and beer, four customers also buy mustard). From the point of view of the analyst, sets \( X \) and \( X' \), and rule \( R \) as well, may be interpreted and validated as knowledge units extracted from the data.

2.2. Data mining methods

The extraction process is based on data mining methods that return knowledge units from the data examined. Data mining methods can be either symbolic or numerical:

- Symbolic methods include, among others: classification based on decision trees, lattice-based classification, frequent itemset search and association rule extraction, classification based on rough sets [Pawlak (1991)], learning methods, e.g., induction, instance-based learning, and explanation-based learning [Mitchell (1997), Michalski et al. (1998)], and database methods based on information retrieval and query answering.
- Numerical methods include, among others: statistics and data analysis, hidden Markov models of order 1 and 2 (initially designed for pattern recognition), Bayesian networks, neural networks, and genetic algorithms.

These methods depend on research domains to which the KDD process is linked [Mannila (1997)]:
• **Statistics and data analysis**: the goal is similar, but the KDD process usually requires a combination of different methods, both symbolic and numerical, as well as domain knowledge for the interpretation of the extracted units.

• **Database management**: database management system techniques may be used to help the data mining task, e.g., using the query capabilities to prepare the data to be mined.

• **Machine learning**: machine learning methods are the core of the KDD process, but scalability, i.e., the amount of data considered, and the objectives are different, i.e., reusing the results of the KDD process for problem-solving or decision-making.

• **Knowledge representation and reasoning**: the data mining process may be guided by a model – a domain ontology – for interpretation and problem-solving.

The KDD process may be considered as a kind of “supervised learning process” since an analyst is in charge of controlling and guiding the process. The analyst may take advantage of his or her own knowledge and of domain ontologies to interpret and validate the results. In this way, the results of the KDD process may be reused to expand existing ontologies, showing that knowledge representation and KDD are complementary processes: *data mining cannot be done without knowledge of the data domain!*

In the following sections, we will primarily be interested in symbolic KDD methods based on the *classification* operation, more specifically, on lattice-based classification, frequent itemset search, and association rule extraction. We will show how the whole transformation process from rough data into knowledge units is based on the underlying idea of *classification*. Classification is a polymorphic procedure involved in every step of the KDD process: within the mining process and the modeling of the domain to design a domain ontology, and within domain knowledge representation and reasoning as well.

### 3. Lattice-based classification

A number of classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), and a binary correspondence between the two classes, indicating, for each individual–property pair, whether the property applies to the individual or not [Barbut and Monjardet (1970), Guénoche and Mechelen (1993), Ganter and Wille (1999)]. The properties may be features that are present or absent, or the values of a property that have been dichotomized as Boolean variables. These variables are collected in Boolean tables relating a set of individuals with a set of properties, where \((i, j) = 1\) or is *true* whenever the individual \(i\) has the property \(j\) (as illustrated in Figure 2).

Lattice-based classification relies on the analysis of such Boolean tables and may be considered as a symbolic data mining technique that can be used to extract from a database a set of concepts organized within a hierarchy (i.e., a partial ordering); frequent itemsets, i.e., sets of properties or features of data that occur together with a certain frequency; and association rules with a given confidence, emphasizing correlations between sets of properties.
More specifically, a lattice is an ordered set \((E, \sqsubseteq)\), where \(\sqsubseteq\) denotes a partial ordering such that every pair of elements \((x, y)\) has an upper bound \(x \lor y\) and a lower bound \(x \land y\) [Davey and Priestley (1990)]. The power-set \(2^E\) of a set \(E\) equipped with the inclusion relation is a basic example of a lattice. The set of natural numbers \(N\) equipped with the divisibility relation is also a lattice: \(x \sqsubseteq y\) if and only if \(y\) is a divisor of \(x\) in \(N\).

A lattice may be built according to the so-called Galois correspondence, classifying within a formal concept a set of individuals, i.e., the extension of the concept, that share a set of properties, i.e., the intension of the concept. Considering the Boolean correspondence between individuals and properties (as shown in Figure 2), it is possible to derive for each individual \(i\) the set of all properties that apply to \(i\). Similarly, it is possible to derive for each property \(j\) the set of all individuals to which \(j\) applies. One may further derive rectangles, i.e., pairs \(O \times A\) where \(O\) is a set of individuals and \(A\) a set of properties, such that every property of \(A\) applies to every individual in \(O\). Moreover, maximal rectangles \(O \times A\) are such that the property set \(A\) consists of all properties the individuals in \(O\) have in common, and the individual set \(O\) consists of all individuals to which the properties of \(A\) jointly apply. Maximal rectangles are called formal concepts: they are concepts because they actually represent a class of objects, where the individual set \(O\) is the extension of the class, and the property set \(A\) is the intension of the class; they are formal concepts because they are mathematical entities that do not necessarily refer to any reality.

From a mathematical point of view, let \(E\) and \(F\) be two finite sets, and \(R\) a binary correspondence on \(E \times F\).

**Definition 1.** The mapping \(f : E \rightarrow F\) is such that, if \(x\) is an element of \(E\), \(f(\{x\})\) consists of all the elements of \(F\) related to \(x\) by \(R\). If \(X\) is an arbitrary part of \(E\), \(f(X) = \{y \in F/\forall x \in X: xRy\}\).

Likewise, the mapping \(g : F \rightarrow E\) is such that, if \(y\) is an element of \(F\), \(g(\{y\})\) consists of all the elements of \(E\) that are related to \(y\) by \(R\). If \(Y\) is an arbitrary part of \(F\), \(g(Y) = \{x \in E/\forall y \in Y: xRy\}\).

The pair \(\{f, g\}\) is said to constitute a Galois connection or a Galois correspondence between the sets \(E\) and \(F\).

In terms of objects and attributes, \(f(X)\) is the set of all attributes shared by all objects in \(X\), and \(g(Y)\) is the set of all objects that have all the attributes of \(Y\). Moreover, \(X \subseteq X' \Rightarrow f(X') \subseteq f(X)\), and \(Y \subseteq Y' \Rightarrow g(Y') \subseteq g(Y)\): the mappings \(f\) and \(g\) are decreasing. For example, considering the Boolean table in Figure 3, we have \(f(\{O_1\}) = \{b,c,e\}\) and \(g(\{b, c, e\}) = \{O_1\}\), \(f(\{O_1, O_2\}) = \{c\}\) and \(g(\{c\}) = \{O_1, O_2, O_3, O_5, O_6\}\), \(g(\{a, c\}) = \{O_2, O_3, O_5, O_6\}\) and \(f(\{O_2, O_3, O_5, O_6\}) = \{a, c, d\}\).

The mapping \(h = g \circ f = g[f]\) maps every part of \(E\) onto a part of \(E\), and the mapping \(h' = f \circ g = f[g]\) maps every part of \(F\) onto a part of \(F\). A subset \(X\) of \(E\) (or alternatively, \(Y\) of \(F\)) is said to be closed if and only if \(X = h(X)\) (or \(Y = h'(Y)\) in the corresponding case).
Let $L_E$ and $L_F$ be the sets of all closed parts of $E$ and $F$, respectively, partially ordered by set inclusion. We may now consider the set $L$ of all pairs of corresponding parts of $L_E$ and $L_F$, i.e., each element of $L$ is the Cartesian product of the closed parts of $E$ and $F$, denoted either by $(X, f(X))$, or by $(g(Y), Y)$, with $X, f(X), Y,$ and $g(Y)$ being closed. The partial order relation $\subseteq$ may be defined on $L$ such that $(X, Y) \subseteq (X', Y')$ if and only if $X \subseteq X'$ (or $Y' \subseteq Y$). The structure $(L, \subseteq)$ is the Galois lattice or the concept lattice of the relation $R$ on $E \times F$, and it can be demonstrated that the elements of $L$ are the formal concepts derived from the relation $R$. For example, the Galois lattice associated with the formal context introduced in Figure 3 is shown in Figure 4.

There are a number of algorithms for building Galois lattices with different specific characteristics [see, for example, Guénoche (1990), Duquenne (1999), Ganter and Wille (1999), Kuznetsov and Obiedkov (2002)]. The notion of the Galois lattice has given rise to the so-called lattice-based classification, and to the active research domain of formal concept analysis² [Ganter and Wille (1999)].

Formal concept analysis is used for a number of different tasks, including the design of object hierarchies, especially class hierarchies in object-oriented programming. Furthermore, lattice-based classification may be used for a number of purposes in KDD [Wille (2002), Valtchev, Missaoui and Godin (2004)].

4. Frequent itemset search and association rule extraction

In parallel with lattice-based classification, one may extract frequent itemsets and association rules from data (as shown in the introductory example in Section 2).

---

² fca-list@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de/mailman/listinfo/fca-list
The extraction of frequent itemsets consists of extracting from formal Boolean contexts sets of properties that occur with a support (as defined below), i.e., the number of individuals sharing the properties, which must be greater than a given threshold. From the frequent itemsets, it is then possible to generate association rules of the form $A \rightarrow B$ relating a subset of properties $A$ with a subset of properties $B$, which can be interpreted as follows: the individuals including $A$ also include $B$ with a certain support and a certain confidence. The numbers of itemsets and rules that can be extracted from a formal Boolean context may be very large, and thus there is a need to prune the sets of itemsets and extracted rules to ensure the subsequent interpretation of the extracted units. This is especially true when the interpretation has to be done by an analyst who is in charge of interpreting the results of the KDD process.

In the following, we introduce the principles governing frequent itemset search and the extraction of association rules. Then practical examples of both processes are proposed.

4.1. Frequent itemset search

**Definition 2.** Given a set of objects $O$ and a set of properties $P$, an item corresponds to a property of an object, and an itemset, or pattern, to a set of items: an object is said to own an item. The number of items in an itemset determines the length of the itemset. The image of an itemset corresponds to the set of objects owning the item.
The support of an itemset corresponds to the proportion of objects that own the itemset, with respect to the whole population of objects. An itemset is said to be frequent if its support is greater than a given frequency threshold $\sigma$: the proportion of objects owning all items included in the itemset is greater than $\sigma$.

For example, let us consider the formal context introduced in Figure 3. If $\sigma = 3/6$, we have: \{a\} is a frequent itemset of length 1 and of support 5/6; \{ac\} is an itemset of length 2, support 4/6, and frequent; \{abc\} has length 3, support 2/6, and is not frequent; \{abcde\} has length 5, support 0/6, and is not frequent. Note that support is a monotonically decreasing function, with respect to the length of an itemset.

When the number of properties in $P$ is equal to $n$, the number of potential itemsets is equal to $2^n$ (actually, the number of all possible subsets of the set $P$): thus, a direct search for frequent itemsets by testing the itemsets that are frequent is not conceivable. Heuristics must be used to prune the set of all itemsets to be tested. This is the purpose of the so-called level-wise search of frequent itemsets, and the well-known Apriori algorithm [Agrawal et al. (1996)] associated with it. The Apriori algorithm relies on two fundamental principles: (i) every sub-itemset of a frequent itemset is a frequent itemset; (ii) every super-itemset of a non-frequent itemset is non-frequent. The Apriori algorithm can be summarized as follows:

1. The search for frequent itemsets begins with the search for frequent itemsets of length 1.
2. The frequent itemsets are recorded and combined together to form the candidate itemsets of greater length. The non-frequent itemsets are discarded; consequently, all of their super-itemsets are discarded as well. The candidate itemsets are then tested, and the process continues in the same way, until no more candidates can be formed.

For example, considering the formal context in Figure 3, with $\sigma = 2/6$, the frequent itemsets of length 1 are \{a\} (5/6), \{b\} (3/6), \{c\} (5/6), and \{d\} (5/6). The itemset \{e\} (1/6) is not frequent and is pruned. Then the candidates of length 2 are formed, combining the frequent itemsets of length 1 – e.g., \{ab\}, \{ac\}, \{ad\}, etc – and then tested. The frequent itemsets of length 2 are \{ab\} (2/6), \{ac\} (4/6), \{ad\} (5/6), \{bc\} (3/6), \{bd\} (2/6), and \{cd\} (4/6). Then candidates of length 3 are formed and tested: the frequent itemsets of length 3 are \{abc\} (2/6), \{abd\} (2/6), \{acd\} (4/6), and \{bcd\} (2/6). Finally, the candidate of length 4, i.e., \{abcd\}, is formed, tested and recorded as a frequent itemset (\{abcd\}(2/6)). No other candidate can be formed, and the algorithm terminates.

When there are huge amounts of data to be mined, i.e., millions of rows and thousands of columns, there is a need to minimize the access to the data to calculate the support. A number of studies have been carried out in this regard, giving rise to very efficient algorithms [see, for example, Pasquier et al. (1999a,b), Stumme et al. (2002)]. Lattices and itemsets are related: actually, the search for frequent itemsets corresponds to a breadth-first search in the concept lattice associated with the formal context under study.
4.2. Association rule extraction

**Definition 3.** An association rule has the form \( A \rightarrow B \), where \( A \) and \( B \) are two itemsets. The support of the rule \( A \rightarrow B \) is defined as the support of the itemset \( A \cup B \) (where \( \cup \) denotes the union of itemsets). The confidence of a rule \( A \rightarrow B \) is defined as the quotient \( \text{support}(A \cup B)/\text{support}(A) \). The confidence can be seen as a conditional probability \( P(B/A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} \), i.e., probability of \( B \) given that one knows \( A \).

A rule is said to be valid if its confidence is greater than a confidence threshold \( \sigma_c \), and its support is greater than the frequency threshold for itemsets, \( \sigma_s \) (a valid rule can only be extracted from a frequent itemset). A rule is said to be exact if its confidence is 1, i.e., \( \text{support}(A \cup B) = \text{support}(A) \); otherwise, the rule is partial.

For example, with \( \sigma_s = 3/6 \) and \( \sigma_c = 3/5 \), \( \{ac\} \) is frequent, and the rule \( a \rightarrow c \) is valid (with support 4/6 and confidence 4/5); the rule \( c \rightarrow a \) is valid (with support 4/6 and confidence 4/5). With \( \sigma_s = 2/6 \) and \( \sigma_c = 3/5 \), \( \{abd\} \) is frequent, the rule \( b \rightarrow ad \) is valid (with support 2/6 and confidence 2/3); however, the rule \( ad \rightarrow b \) is not valid (with support 2/6 and confidence 2/5).

The generation of valid association rules from a frequent itemset (of a length necessarily greater than or equal to 2) proceeds in a similar way to the search for frequent itemsets. Given a frequent itemset \( P \), the extraction starts by generating the valid rules with a conclusion of length 1, say, rules of the form \( P \setminus \{i\} \rightarrow \{i\} \), where \( i \) is an item of length 1, and \( P \setminus \{i\} \) denotes the itemset \( P \) without the item \( \{i\} \). Then, the conclusions of the valid rules \( P \setminus \{i\} \rightarrow \{i\} \) are combined to generate the candidate conclusions of length 2, e.g., \( P \setminus \{ij\} \rightarrow \{ij\} \), and the process continues until no more valid rules can be generated from the frequent itemset.

For example, with our current formal context, given \( \sigma_s = 2/6 \) and \( \sigma_c = 2/5 \), when \( P = \{ab\} \), the valid rules generated are \( \{a\} \rightarrow \{b\} \) (2/6,2/5) and \( \{b\} \rightarrow \{a\} \) (2/6,2/3). Given the frequent itemset \( P = \{abc\} \) (2/6), the generated rules are \( \{ab\} \rightarrow \{c\} \) (2/6,1), \( \{ac\} \rightarrow \{b\} \) (2/6,1/2), and \( \{bc\} \rightarrow \{a\} \) (2/6,2/3). Since \( \{a, b, c\} \) are three valid conclusions, they can be combined to produce the new conclusions \( \{ab, ac, bc\} \), and generate the rules \( \{c\} \rightarrow \{ab\} \) (2/6,2/5), \( \{b\} \rightarrow \{ac\} \) (2/6,2/3), and \( \{a\} \rightarrow \{bc\} \) (2/6,2/5), which are all valid rules.

5. Applications

In the following subsections, we shall present three applications of the KDD process that rely on the data mining techniques presented above: an experiment in mining reaction databases for organic chemistry planning, an application in mining gene expression databases in biology, and an introduction to Web mining that concludes the section.
5.1. Mining chemical reaction database

In this subsection, we shall examine an experiment on the application of knowledge discovery algorithms for mining databases of chemical reactions [Berasaluce et al. (2004a, b)]. Chemical reactions are the main elements on which synthesis in organic chemistry relies, and this is why chemical reaction databases are of the first importance. Synthesis planning is mainly based on retrosynthesis, i.e., a goal-directed problem-solving approach, where the target molecule is iteratively transformed by applying reactions to obtain simpler fragments, until accessible starting materials are found (see Figure 5). This experiment was designed to discover generic reactions, also called synthesis methods, from chemical reaction databases in order to design generic, reusable synthesis plans.

5.1.1. The chemical context

Two main categories of reactions may be distinguished: reactions that build the skeleton of a molecule – the arrangement of carbon atoms on which any organic molecule relies – and reactions that change the functionality of a molecule, i.e., change one function into another function. We will mainly be interested in reactions that affect functionality, and especially in the following question: what reactions allow the transformation of a function $F_i$ into a function $F_j$?

The information items in reaction databases such as ORGSYN-2000 (Organic Syntheses database, which contains 5486 records) and JSM-2002 (Journal of Synthetic Methods database, which includes 75,291 records) may be seen as a collection of records, where every record contains one chemical equation involving structural information, which can be read, according to the reaction model, as the transformation of an initial state – or the set of reactants – into a final state – or the set of products – associated with an atom-to-atom mapping between the initial and final states (see Figure 6).

The transformations considered are functional modifications and functional addition and deletion, i.e., adding or deleting a function. The reactions were considered at an abstract level, the so-called block level, as shown in Figure 7.

Fig. 5. The general schema of a synthesis problem.
5.1.2. Mining of a reaction database

The RESYN–ASSISTANT system was used to recognize the building blocks of reactions [Berasaluce et al. (2004a,b)]. Based on atom-to-atom mapping, the system establishes the correspondence between recognized blocks of the same nature, and determines their role in the reaction. A function may be present in a reactant, a product, or both. In the latter case, the function is unchanged. In the other two cases, the function in the reactant is destroyed, or the function in the product is formed. During a reaction, one or more reactant functions may contribute to form the product functions. At the end of the preprocessing step, the information obtained by the recognition process is incorporated into the representation of the reaction.

In order to allow the application of the algorithms for frequent itemset search and association rule extraction, namely the Close algorithm [Pasquier et al. (1999a,b)], the data on reactions were transformed into a Boolean table (thereby losing the actual representation of a molecule as a composition of functional blocks). A reaction can then be considered from two main points of view (see Figure 8):

- A global point of view on the functionality interchanges, leading one to consider a single entry $R$, corresponding to a single analyzed reaction, with which a list of properties, i.e., formed and/or destroyed and/or unchanged functions, is associated,
- A specific point of view on the functionality transformations that is based on the consideration of two (or more) different entries $R_k$ corresponding to the different functions being formed.
Both points of view were examined during the experiment. The Close algorithm was applied to Boolean tables to generate first itemsets, i.e., sets of functions (with an associated support), and then association rules. The frequent itemsets extracted may be studied from different points of view. Studying frequent itemsets of length 2 or 3 enables the analyst to determine basic relations between functions. For example, searching for a formed function $F_f$ (f for formed) derived from a destroyed function $F_d$ (d for destroyed) leads to the study of the itemsets $F_d \cap F_f$, where the symbol $\cap$ stands for the conjunction of items or functions. In some cases, a reaction may depend on functions present in both the reactants and the products that remain unchanged ($u$ for unchanged) during the reaction application, leading to the study of frequent itemsets such as $F_f \cap F_u \cap F_d$. This kind of itemset can be searched for and analyzed to extract a “protection function” that is supposed to be stable under the given experimental conditions.

The extraction of association rules gives one a complementary perspective on the knowledge extraction process. For example, searching for the most frequent ways of forming a function $F_f$ from a function $F_d$ leads to the study of rules such as $F_f \rightarrow F_d$; indeed, this rule must be read in a retrosynthetic way, i.e., if the function $F_f$ is formed, this means that the function $F_d$ is destroyed.

5.1.3. Discussion

A qualitative and statistical study of the results has shown the following behaviors. Some functions have a high stability, i.e., they mostly remain unchanged, while, on the contrary, other functions are very reactive, i.e., they are generally destroyed. All the reactive functions are more present in reactants than in products, and some functions
are formed particularly often. Some functions, which are among the most widely used in organic synthesis, are most often present and destroyed in reactants, e.g., alcohol and carboxylic acid.

Knowledge is used at every step of the knowledge extraction process, e.g., the coupling of the knowledge extraction process with the RESYN–ASSISTANT system, and domain ontologies such as the function ontologies. Indeed – and this is one of the major lessons of this experiment – the knowledge discovery process in a specific domain such as organic synthesis has to be knowledge-intensive, and has to be guided by domain knowledge, and by an analyst, for obtaining meaningful results.

Moreover, the use of data mining methods such as frequent itemset search or association rule extraction proved to be useful, and provided encouraging results. It would be interesting to test other (symbolic) data mining methods, especially relational mining, in order to be able to take the structure of molecules into account for the data mining task [Dzeroski and Lavrac (2001)].

5.2. An experiment in biology

In this subsection, we will present an experiment on the mining of gene expression databases to extract association rules, based on Creighton and Hanash (2003) [see also Wang et al. (2005) for a recent overview of data mining in bioinformatics]. Global gene expression profiling can be a valuable tool in the understanding of genes, biological networks, and cellular states. One goal in analyzing expression data is to try to determine how the expression of any particular gene might affect the expression of other genes; the genes involved in this case could belong to the same biological network.

As larger and larger gene expression data sets become available, data mining techniques can be applied to identify patterns of interest in the data. Creighton and Hanash (2003) describe an experiment where an Apriori algorithm was applied to mine association rules from gene expression data, using a set of 300 expression profiles for yeast. An example of an extracted association rule is the following: \{cancer\} \rightarrow \{gene A \uparrow, gene B \downarrow, gene C \uparrow\}, meaning that, for the data set that was mined, in most profile experiments where the cells used were cancerous, gene A was measured as being up (highly expressed), gene B was down (low expression), and gene C was up. In the context of formal databases, a gene expression profile can be thought of a single transaction (corresponding to a row in a Boolean table), and each protein can be thought of as an item.

The extracted association rules that were considered in the experiment are of the form \{LHS\} \rightarrow \{RHS\}, where \{LHS\}, i.e., left-hand side, is composed of only one item, and \{RHS\}, i.e., right-hand side, may have an arbitrary number of items. Note that such association rules, where \{LHS\} is composed of only one item, are very interesting and are the basis of efficient algorithms for a level-wise itemset search [Stumme et al. (2002)]. Actually, in Creighton and Hanash (2003), closed itemsets were considered, and the set of extracted association rules was manually pruned to make the results more understandable. In particular, this shows the importance of presenting small sets of association rules or frequent itemsets for the indispensable human analysis of the results.
Two examples of extracted association rules are given below. The minimum support was set at 10%, and the minimum confidence at 80%, and a rule may be interpreted as follows: when the gene in \{LHS\} is up, so are the genes in \{RHS\}. The expressions of rules have been simplified to ensure better readability for non-biologists.

\[
\begin{align*}
\{YHM1\} &\rightarrow \{SEQ1, ARO3\} \\
\{ARO3\} &\rightarrow \{SEQ1, YHM1\}
\end{align*}
\]

where \{SEQ1\} = \{ARG1, ARG4, CTF13, HIS5, LYS1, RIB5, SNO1, SNZ1, YHRO29C, YOL118C\}

The genes \{YHM1\} in the first rule and \{ARO3\} in the second rule are found on opposite sides of the rules. The gene \{YHM1\} has been identified as a suppressor of a gene that has the property of being a binding factor. On the other hand, the gene \{ARO3\} is activated by a binding factor itself. Whether the nature of the association between \{ARO3\} and \{YHM1\} suggested here has something to do with the fact that both of these genes have an association with a binding factor is an open – and very interesting – question.

The association rules that have been mined represent only a fraction of all the possible gene-to-gene interactions that remain to be discovered in yeast. More rules can be found using different search criteria, i.e., changing the support, the confidence, the data, and the form of the extracted rules. The extracted association rules can lead to the generation of new hypotheses explaining some aspects of the gene interactions, to be confirmed in wet laboratory experiments. Mining expression data for association rule extraction seems to be most useful for interpreting and understanding gene networks: association rules can describe how the expression of one gene may be associated with the expression of a set of genes. It should be noted that an association rule implies an “association,” which is not necessarily a “cause and effect” relationship. Determining the precise nature of the association requires expert knowledge, as emphasized in the subsection on the mining of chemical reaction databases. This study shows that it is possible to develop bioinformatics applications that do more than merely storing and retrieving expression data, and to propose tools for exploratory data analysis.

5.3. An introduction to Web mining

In the framework of the Semantic Web, machines talk to machines so that they can deliver services to people [Fensel et al. (2003)]. Tomorrow the Web will be a distributed, shared, declarative, and navigable space; it will be mainly exploited by computers that solve problems for humans and give the results of their operations to humans. The semantics of documents on the Web must therefore be accessible to computers. One important element of this semantics is an explicit model of the domain of data, describing the vocabulary and the structure of information in relation to the domain of interest. This model must be commonly accepted and shared: this is the essence of the notion of ontology, as it applies in the Semantic Web framework and is used to build knowledge systems. For example, let us consider the following list of queries, which raise a series of different, and complex, problems:
• A book on Félix Leclerc.
• A book (written) by Félix Leclerc or a book by Félix Leclerc.
• A biography of Félix Leclerc.
• An autobiography of Félix Leclerc.
• A songbook by Félix Leclerc.
• A book on the work of Félix Leclerc.

To respond to these queries, a computer system has to understand the actual meaning of the questions (the user’s “intended meaning”); it must also be able to make the distinction between on in a book on and of in a book of, and to understand the difference between terms such as book, songbook, biography, autobiography, etc. This is the purpose of ontologies in the Semantic Web and Web mining framework [Staab and Studer (2004)]. Moreover, it may also be very useful for the system to know who Félix Leclerc is in order to respond to the above queries (as it would be for a human himself as well).

The description of the content of documents may be made explicit by using document description languages such as XML. A semantics can be attached to documents – and their content – using knowledge representation languages, e.g., description logics and Ontology Web Language (OWL) [Fensel et al. (2003)]. An intelligent manipulation of documents is based on the exploitation of their content and semantics, with respect to the knowledge of the domain of the documents. The technology for the Semantic Web is based, on the one hand, on the use of languages for ontology representation, and for document and resource description such as XML and Resource Description Framework Schema (RDF (S)) (and on the other hand, on the use of intelligent search engines and mining modules to improve the retrieval of adequate resources for problem solving. In this way, information extraction – extraction of key terms from documents – and data mining – especially text mining – may be used to analyze and classify documents with respect to their content [Carpinetto and Romano (2004) may be of interest regarding the content-based information retrieval and lattice-based classification of documents].

The mining of documents on the Web, or Web mining, can be carried out with three main goals [Kosala and Blockeel (2000), Berendt, Hotho and Stumme (2002)]:

• The mining of the content of documents, in relation to text mining [see Chapter 42, page 935 (this volume) and Janetzko et al. (2004), for example].
• The mining of the structure of the pages and of the links between pages (hypertext links).
• The mining of usages or sets of operations applied on pages.

Web mining can be a major technique in the design of the Semantic Web: on that basis, ontologies can be designed semi-automatically, rather than manually. Ontologies can be used to annotate documents, and thus to enhance the document mining process, on the basis of document content. The Web mining process can be used to improve the annotation of documents, and thus the semantics attached to them, i.e., their content, understandability, and structure.
6. Discussion

The KDD process must be carried out in a KDD environment where data mining is guided by domain knowledge and embedded in ontologies and knowledge-based systems. The knowledge units used in knowledge systems may have two main sources: explicit knowledge, which can be contributed by domain experts, and implicit knowledge, which must be extracted from databases of different kinds, e.g., rough data or textual documents. In addition, an important issue in the framework of the Semantic Web and Web mining is the ability to manipulate documents by their content, in order to search, annotate, and classify them. The content-based manipulation of documents allows one to solve a number of problems such as information extraction, intelligent information retrieval, content-based document mining, and more. More specifically, Brachman and Anand (1996) list the following requirements for knowledge discovery tools:

- The system should represent the underlying domain and present it to the user in a natural and appropriate fashion. Objects of the domain should be easily incorporated into queries.
- The domain representation should be extendible by the addition of new concepts or classes formed from queries. These concepts and their representative individuals must be usable in subsequent queries.
- It should be easy to form tentative data segmentations, to investigate the segments, and to resegment quickly and easily. There should be a powerful repertoire of viewing and analysis methods, and these methods should be applicable to segments (as in the Weka system, for example [Witten and Franck (2000)]).
- Analysts should be supported in recognizing and abstracting common analysis (segmenting and viewing) patterns. These patterns must be easy to apply and modify.
- There should be facilities to monitor changes in classes or concepts over time.
- The system should increase the transparency of the knowledge discovery process and should document its different stages.
- Analysis tools should take advantage of the explicitly represented background knowledge of domain experts, but should also activate their implicit knowledge.
- The system should allow highly flexible processes of knowledge discovery respecting the open and procedural nature of productive human thinking. This means, in particular, supporting intersubjective communication and argumentation.

7. Conclusion

The knowledge discovery in databases process consists of processing a huge volume of data in order to extract knowledge units that can be reused either by an expert in the domain or by a knowledge-based system to solve problems in the domain. The KDD process is based on three major steps: data preparation, data mining, and interpretation of the extracted units. Moreover, the KDD process is iterative and interactive, and is
controlled by an analyst, who is in charge of guiding and validating the extraction process. In addition, the KDD process may take advantage of domain knowledge, i.e., ontologies and knowledge bases, to improve every step of the process. Data mining methods are divided into two main categories: symbolic and numerical. In this chapter, we focused primarily on symbolic methods, and especially on lattice-based classification, level-wise search for frequent itemsets, and association rule extraction. These methods are operational and can provide good results in real-world problems. Indeed, three kinds of applications were described: an experiment on the mining of chemical reaction databases, an experiment on the mining of gene expression databases, and finally, an increasingly important research field, Web mining.

Regarding the future of KDD, many problems remain to be solved, at every step of the process. Two important issues are related to the KDD process, firstly as a knowledge-guided process, as we have tried to show, and secondly as a complete system integrating knowledge representation and database management functionalities. Another important area for the investigation of symbolic methods is the extension to the processing of complex data (in contrast to Boolean data). Finally, let us mention that some key challenges are linked to the application domains, and must still be addressed, e.g., biology, chemistry, medicine, space, weather forecasting, finance, etc. At the beginning of this chapter, we compared knowledge discovery to panning for gold research or doing an archaeological dig: first, one needs to become familiar with the domain of the data, then to apply a number of data mining methods that may produce more or less useful results, and then to validate these results. Meanwhile, the analyst must be patient because the process is iterative – the work is long and may never be successful – but it is definitely worthwhile continuing the job, and remaining confident and optimistic!

References


Michalski, R., I. Bratko and M. Kubat, eds (1998), Machine Learning and Data Mining (Wiley, Chichester, UK).


Witten, I., and E. Frank (2000), Data Mining (Morgan Kaufmann Publishers, San Francisco).
This page intentionally left blank
Chapter 42

GENRE-SPECIFIC TEXT MINING AND EXTENSIONAL INDUCTIVE CONCEPT RECOGNITION:
A PSEUDOCOGNITIVE APPROACH

YVES KODRATOFF

CNRS, U. Paris-Sud

Rúnar munt þú finna ok ráðna stafi …
Runes will thou find and good-counseling engraved-letters …
(Hávamál, verse 142)

Contents

Abstract 936

1. Introduction and definition of text mining (TM) 937
   1.1. Text mining 937
   1.2. Our approach 938

2. Text retrieval 939

3. Standardization 940

4. Grammatical tagging 940
   4.1. Why expert rules at the tagging stage? 941
       4.1.1. The accuracy rates are correct, but this is still not “good enough” 941
       4.1.2. Accuracy on what? 941
       4.1.3. Recursion 941
   4.2. A tagging language 942
   4.3. Our approach to grammatical tagging 943
   4.4. Automatic learning of new tagging rules 943

5. Terminology 944

6. Concept recognition in texts 944
   6.1. Polysemy 945
   6.2. General versus local collocations 945
   6.3. Terms and collocations 945
   6.4. ACT as a friendly interface helping the expert 946
   6.5. ACT as an inductive program 947
   6.6. Validation 951
       6.6.1. Validation of the concepts 951
       6.6.2. Validation of the induction 951

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved

935
Abstract

This chapter describes a software tool that is being developed to help domain experts design procedures in order to be able to recognize the linguistic instances of a set of concepts in texts relative to their field. This involves several challenging steps, from cleaning to concept tagging. The problem is made more difficult by the fact that these steps are not independent, but are recursively embedded in each other.

In this chapter, two crucial notions are explored. One is that only a domain expert can develop tools able to solve these problems. It follows that the computer scientist should develop user-friendly tools enabling the expert to transfer the expertise to the programs. The second is that inductive tools must also be provided, or the workload will be so heavy that nothing substantial can ever be achieved. The difficulty is that induction has to take place from data that are both incomplete and very noisy; this is a well-known cause of failure in most of the existing inductive programs. Special inductive procedures, of various kinds, have to be introduced in the process. In particular, in order to solve the difficult problem of concept recognition in texts, that is, gathering a large number of possible instances of each concept in texts, a new form of learning has been defined, called Extensional Induction.
1. Introduction and definition of text mining (TM)

1.1. Text mining

The purpose of this chapter is to describe how concept recognition can be learned inductively or can be achieved by inductive learning from texts and examples provided by a specialist, in order to develop efficient procedures to perform “TM,” as defined below. This approach is dubbed pseudocognitive because there is no attempt to try to define what a concept is; rather, the goal is to recognize its presence in a text.

As a consequence of this extensional approach, the domain expert is required to provide several examples of what he or she identifies as a concept in a text, that is, the linguistic instances of the concept’s presence, and we try to inductively and automatically increase the number of recognized linguistic instances. This is why it is more apt to speak of “learning in extension,” by increasing the number of linguistic instances of a concept. Similarly, it is better to use the term “concept recognition,” rather than “concept learning,” which is what happens when “learning in intention” takes place; in other words, when what is sought are the feature-values defining the concept.

The following definition of TM underlines the fact that the goal of TM is not to understand the text but to use it to achieve a specific task:

**TM is a set of approaches that perform some kind of transformation of a text in order to extract its meaning with a view to improving the solution of a specific problem.**

The difficulty of this domain lies in the delicate balance that must be maintained between the texts, the transformations, and the problems.

Problem solving implies the existence of cognitive entities that can be called field-specific concepts: they are part of the knowledge necessary to resolve the “specific problem” TM is addressing. Our basic hypothesis calls for a minimal theorization of the language, but one that is sufficient to implement algorithms inducing field-specific concepts and cause these concepts to be recognized in texts by means of some linguistic instance of the concept presence. For example, in the human resources texts we study, the verb to give up, followed by the noun influence evokes the presence of a concept that psychologists recognize as describing the nature of relationships within the company, which they call “relational.”

The citation from the Hávamál above illustrates our point very accurately. The Rune God, known as Odin or Woden, claims that a few scrapings (staff) on a stone are not yet a physical instance of the concept of runes. These scrapings have to be of good counsel to acquire runic status. Runologists tend to recognize physical instances of the rune concept in a very different way, leading to an enormous shift in the definition of the concept. When it comes to more practical problems, our basic hypothesis is that, depending on the problem, domain specialists tend to slightly shift the definitions of the concepts they use, and these shifts are noticeable in the linguistic instances produced when they speak of these concepts.
1.2. Our approach

They are many kinds of linguistic instances, and we started with the strong working hypothesis (which is expected to be revised) that the linguistic instances relevant to concept recognition are collocations. For a linguistically sound definition of collocations, see the works of Halliday [e.g., Halliday and Matthiessen (2004)].

A nominal lexical dependency consists of two entities: a pivot and a term. A term is a nominal expression that contains a noun associated with one or more adjectives, adverbs, or other nouns. By this definition, a term never contains a verbal group. The pivot can be a verb or verbal group, an adverb, a noun, an adjective, or another nominal term. When one of the two components of the nominal dependency contains a verbal group, or does not contain a noun, then it is the pivot. Otherwise, if both components are nominal terms, then the user has the choice between “the first is the pivot” and “the second is the pivot.” The first is the default option in French, the second in English.

Many statistical or linguistic treatments are needed before any kind of information can be extracted from a text. Instead of focusing on one specific step, we insist that it is necessary to deal with the complete chain of steps, especially since we take into account the links between successive steps. A chart describing the field of TM is presented in Figure 1. Our approach is indicated by the solid lines in this figure.

Since we are dealing with massive amounts of texts, all these steps have to be automated, including their mutual interactions. It follows that no step is looked at in isolation. It might even happen that a “high-level” step is used to improve a “low-level” one. For instance, terminology is a well-developed field, and we do not seek much improvement in terminology itself; rather, we want to improve tagging in order to obtain better terminology – better for conceptual categorization and writing extraction patterns.

There is yet another constraint due to the fact that we are interested in texts written for specific purposes. Domain experts play an important role, and thus user-friendly software has to be created for them. Since expert time is very expensive, usability is not enough – we must include inductive processes that will speed up the whole process.

We are now building tools to help a domain specialist perform the successive steps shown in Figure 1. The tools we will specifically focus on are those belonging to the following subchain:

standardization → tagging → terminology → concept recognition in texts

This application is difficult since it demands some understanding of the text content, especially when the information is extracted from tables. The exact meanings of the rows and columns must be known. We are also interested in two other applications. One is association detection, and we have developed a technique for finding “interesting” associations with little support (what is called in the data mining (DM) jargon a “nugget”). The other one is the comparison of an existing ontology with the texts dealing with the topic of the ontology. In this chapter, I will describe in detail the steps needed to recognize concepts, but I will supply no details about the possible applications.
2. Text retrieval

Our team is presently working with five technical corpora.

A “biology” corpus was obtained by querying the National Library of Medicine’s Medline (PubMed) database with the key words DNA-binding, proteins, and yeast; the result is a corpus of 6119 abstracts with a volume of some 10 MB. We are presently in the process of gathering some of the corresponding complete papers. This corpus illustrates the problem of dealing with many different texts written in a highly technical language.

A “DM” corpus (369 kB) was obtained from the introductions to papers at Knowledge Discovery and Data (KDD) mining conferences. The “English speakers” subcorpus contains 100 texts authored by people working in English-speaking countries, while the “French speakers” subcorpus contains 31 texts in English authored by native French speakers personally known to the creator of the corpus. This subcorpus illustrates the problem of dealing with texts written in semitechnical language and, as a side-effect, gives some indications of the problems encountered by French speakers when they write in English.
The “human resources” corpus comprises 3784 kB of text belonging to the company PerformanSe. It is written in French by a single psychologist who serves as an expert, and illustrates the problem of mining well-balanced sentences in one style.

A “CV” corpus (2470 kB) contains 1144 CVs written in French, belonging to the company VediorBis. It illustrates the problem of mining texts written in semitelegraphic style with many spelling mistakes.

Finally, there is 40 MB of a “newspaper” corpus, part of the training data provided by the TREC competition. We learned a lot from this last corpus, but since participation in the TREC competition involves signing confidentiality clauses, no details about this corpus will be given.

3. Standardization

Among other steps, standardization contains one most that is overlooked because very few general laws can be formulated for this purpose, but that is of primary importance in view of our definition of TM: text cleaning. This process is obviously specific to each domain, and only an expert is able to decide what should or should not be cleaned. In general, we have noticed that this step influences the proper functioning of all subsequent linguistic processing but it comprises a very large number of highly specific rules. In addition to some seemingly trivial transformations such as suppressing the authors’ names, database formatting, etc., we perform two types of cleaning.

First, the vocabulary used is different and, for example, we replace the various references to the USA by a uniform “United_States,” tagged as a name not designating an individual person. Second, each specialty develops its own jargon, and this jargon must be standardized.

4. Grammatical tagging

The lexicon itself contains hidden semantic information about the possible roles of each word in the text. For instance, *will* and *may* can be modal verbs, nouns, or first names (when capitalized), while *would* can only be a modal, as it always introduces the semantics of a possible action. By choosing among the possible tags, syntactic tagging introduces even more semantic possibilities. For instance, one’s analysis of a sentence will be totally different depending on whether or not one knows whether *will* refers to someone named Will, or a person’s will is expressed, or a modal form is used. Even though the lexicon and grammatical tagging introduce a large amount of information, this still does not completely solve the problem of polysemy.

However, grammatical tagging is however often considered to be field-independent, and we need to address in more detail the issue of why we want to give the expert the possibility to decide on contextual tagging.
4.1. Why expert rules at the tagging stage?

Excellent results have often been reported using a supervised learning technique: inductive techniques using a “hand-tagged” corpus. Some new techniques are even able to reach more than 95% accuracy rates, and it may seem that this is good enough.

4.1.1. The accuracy rates are correct, but this is still not “good enough”

The taggers being trained on human-tagged corpora require an enormous investment, as we have seen in general areas. The available corpora are extremely expensive and taggers are not available in specialized domains where a domain expert is needed. We are thus forced to use taggers trained on general topic areas, and their accuracy drops dramatically when applied to topics they were not trained upon. We are currently in the position of correcting the tags in an already tagged corpus, which is too inaccurate to be used in its present state.

4.1.2. Accuracy on what?

Even when the tagging is very accurate, authors tend to forget to report the nature of the mistakes made by their taggers. For instance, we checked Brill’s (1994) tagger on a corpus where it was fairly accurate [Amrani, Kodratoff and Matte-Tailliez (2004)] but some tags reached an error rate of 100%.

In fact, some errors are much more harmful than some others. For instance, the word that is very ambiguous, and a mistake on its grammatical tagging can destroy the structure of the sentence. Similarly, confusion between nouns and verbs will entail a gross misunderstanding of the sentence. The “tagging community” should develop a register of such harmful errors, and the error rates on these kinds of errors should be specified in papers that claim a high accuracy rate.

4.1.3. Recursion

Some errors occur in doublets or triplets that make their correction especially difficult and their destructive possibilities all the greater. In this case, we “tricked” recursion by using time and side-effects of the transformations. Suppose, for instance, that we encountered the pair alternate/VBP projects/VBZ, meaning that alternate is tagged as a verb and projects as a third-person singular verb. Now suppose we want to correct it to alternate/JJ projects/NNS, meaning that an adjective is followed by a plural noun, we introduce a special tag signaling the origin of the corrected tag, say NNS_exVBZ. We then write two rules. The first one says that, in some contexts, if projects follows alternate/VBP then it has to be retagged NNS_exVBZ. The second one says that, in the same context, if alternate/VBP is followed by either projects/VBZ or projects/NNS_exVBZ, then it has to be tagged as an adjective. Finally, all NNS_exVBZ
tags are transformed back into NNS tags. The actual implementation is a bit more complex, in order to avoid so many specific cases, but it makes use of this very idea.

4.2. A tagging language

All this explains why we developed a language to tag the texts. Since we suppose that we already have a lexicon and tagged texts (both of which certainly contain mistakes), this language deals essentially with the relational tagging problems. It takes the form of rules handling triplets that describe – in this order – the place of the word in the sentence relative to a fixed word of interest placed at “0,” the word itself, and its tag. For instance, in order to avoid writing overly specific rules, we would build up list1 containing alternate and list2 containing projects, and the rules described earlier would be written as follows:

\[
\text{if } [\text{context}] (0, \text{list1}, \text{VBZ}) (1, \text{list2}, \text{NN}) \text{ then } (1, \text{NNN}, \text{exVBZ})
\]

\[
\text{if } [\text{context}] (0, \text{list1}, \text{VBZ}) (1, \text{list2}, \text{orNNN,exVBZ}) \text{ then } (0, \text{JJ}) (1, \text{NN})
\]

This language will be made freely accessible on the Internet as soon as the user’s manual is ready, so I shall not provide too many details about it here. The rules presented above describe the simplest cases. In general, it is designed to write quite complex relationships among words in one sentence, using relatively simple formulae. For instance, to properly tag a that as an IN [define IN], we wrote a rule stating: “tag a that as IN when the first verb before this that belongs to a given list, except when there is a singular or plural noun (not a name) between that and this verb.”

A very important feature of the language is that it easily describes cases, where optional words exist inside a relation. For instance, the premise of a rule describing a determiner followed by a noun is written as:

\[
(1, \text{DT}) (0, \text{NN})
\]

If, between the noun and the determiner, one adverb (called “RB”) or two adjectives with a coordinating element between them are possible but not necessary, we simply write this premise as

\[
(1, \text{DT}) (*1, \text{orJJ, RB}) (*1, \text{JJ}) (*1, \text{CC}) (*1, \text{JJ}) (0, \text{NN}).
\]

Note that this language is nothing but a Perl interpreter for generating Perl programs that will perform the job requested by the formula. It might be believed that a skilled Perl programmer could do this work easily. In fact, since we use the language to check various hypotheses concerning the nature of relationships entailing a tag, we write a very large number of such formulae. Since each Perl program is relatively complex, even a skilled Perl programmer – not to mention the domain expert – might prefer using our language when hundreds of programs are called for. Anyone can learn our language
in a few hours since it is straightforward, and we kept the semantic difficulties down to a total of three, sometimes favoring simplicity over power.

4.3. Our approach to grammatical tagging

Our present methodology is as follows:
1. Clean the corpus and build a lexicon.
2. Use our lexicon and an existing free tagger to perform an initial tagging. In this step, we use Brill’s tagger because it is free, open software able to use any lexicon and it can be retrained very easily as our tagging progressively improves. Another, less determining, reason is that its French version (freely available from CNRS-Nancy) has been very carefully trained and gives excellent results on French general topic texts.
3. Write specialized rules to handle the remaining mistakes. In fact, the only rules we have written up to now cover the case of *that*, confusion between verbs and nouns, confusion between adjectives and verbs, and mistakes that spoil the context in which our rules apply such as most of the cases of confusion between nouns and adjectives. Note that once the confusion between adjectives and past participles is solved, it becomes very easy to recognize passive forms, and we have introduced a few rules to that effect.

We have also designed a user-friendly interface, which shows the results of the rules and can perform simple transformations that do not need our tagging language.

4.4. Automatic learning of new tagging rules

Learning relational laws within a sentence is obviously a relational problem, which in principle demands Inductive Logic Programming (ILP). The usual problem with ILP is that its search space is very large, and learning tends to be very slow. In the case of natural language sentences, so many word combinations are possible that the definition of the set of features (even before one thinks of searching for any solution) is already so large that simply “defining the problem” is more or less impossible. For instance, the limit on the distance from a verb governing a choice between a noun and an adjective tag cannot be known in advance. The sentences are generally under 50 words long, and yet the number of possible combinations of words, tags, and distances within the sentence is still so large that defining a fixed set of features is a hard preliminary learning task.

By asking the experts to generate the sets of rules, we also obtain sets of features the values of which decide on a tag. This is how we “cheated” with the preliminary learning task. We use the sets of expert-defined features – as extracted from the rules each expert wrote – to define the learning search space. In this way, the relations to be sought are known in advance and we can bypass ILP by introducing a few features that are actually functions taking a variable value, depending on the text they are applied to. This enables us to use a propositional learner to learn relations.
A supplementary advantage of our approach is that the rules we will ultimately obtain will be composed of expert-understandable pieces. As already explained, we are not expecting to ever obtain a perfectly tagged corpus; we are simply trying to cut down on mistakes, especially those mistakes the expert considers to be unacceptable. When the expert is really convinced that the rules he/she produces are 100% accurate, the system can “meta-tag” them as such. In this way, we obtain a corpus that is tagged with both unconfirmed tags and exact tags.

In all cases, note that our goal is not to obtain staggering levels of precision, but to obtain expert-understandable rules that can be validated independently of our own corpus.

5. Terminology

Good tagging is very important for our terminology since we have to carefully differentiate between a (nominal) term and a (verbal) pivot. The further induction steps hinge on these notions, so we have to be as accurate as possible in that respect. In addition, and since we build up our terminology with a view to clustering the terms in conceptual classes, the described concept, and its behavior relative to parent concepts, will be very different depending on the nature of the elements of the term. For instance, given an existing term, term1, then the new term built as adjective-term1 is certainly a subterm of term1 since adjective qualifies term1. As for noun-term1, it may be a subterm of term1, but it will more likely be a parent term with a similar but independent meaning.

We constructed a system, named EXIT, enabling the user to set the parameters, observe the results obtained, write rules differentiating between interesting and uninteresting terms, and tag the terms as “rejected” or “accepted,” with a choice of reasons for rejection. In particular, rejections due to errors in the cleaning and tagging are easily remembered, retrieved, and introduced in the subsequent iteration without forcing the user to stop working on terminology. Since our whole process is goal-oriented, it often happens that a perfectly valid term is valid for the domain as a whole, but not for our specific goal. The expert is requested to tag this term as “rejected” with a supplementary tag specifying that it does not fit among the concepts the expert is looking for but could be used in another application.

We should also note that EXIT’s interactivity generates some useful side-effects. Word grouping can be performed for any kind of grammatical category by modifying the regular expressions that select the doublets or triplets to be linked together. Thus, accepting a term amounts to declaring that the sentence flows smoothly over this group of words. Rejecting a term, on the contrary, amounts to cutting the sentence into pieces. This provides a structure for the sentence that can be used in later steps.

6. Concept recognition in texts

Once terminology is completed, we can use it to recognize the occurrence of a concept in a text, by clustering the terms into classes, with each term being an instance of a concept.
We developed a system of concept recognition in texts [Fontaine and Kodratoff (2002), Kodratoff (2001)], called “Acquisition de Concepts à partir de Textes” (ACT). This system is not yet complete, and we are presenting here an interim view of ACT, as it works in June 2004.

6.1. Polysemy

The task of concept recognition is usually seen as intractable, due partly to polysemy. In fact, we find three kinds of polysemy:

- Accidental polysemy occurs when the tagger makes a mistake and confuses two grammatical occurrences of the same string of letters. For instance, the noun plane (a flat surface) may be confused with the related adjective plane. This is not real polysemy but it introduces a similar kind of confusion.
- True polysemy occurs when two identical sequences of letters have totally different meanings, for instance, plane meaning ‘surface’ and plane meaning ‘aircraft.’
- There is also the situation that occurs when the same object plays different roles in the text. This definition is controversial, since it is often said that, for instance, the aircraft plane cannot be polysemic with another kind of plane: their roles are different. It is true that the word plane plays a different role in the sentences the plane bombed the tank and in the plane made mass tourism possible. All the grounded instances of bombers are different from each instance of an airliner; therefore we would claim that plane, as a bomber, and a plane, as a means of passenger transportation, are polysemantic (to be more precise, they are “grounded-polysemic”). We consider that two sequences of letters play different roles only when their grounded instances are at least partially confused. For instance, the same satellite can be seen as playing the role of a flying object or that of an information transmission device.

Note, however, that from our point of view, we look at strings of words as being instances of concepts. In that case, roles and true or false polysemy are essentially the same: the same string of letters points at different concepts in the text, and we have to recognize their difference.

6.2. General versus local collocations

Our approach opposes the so-called “Harrisian hypothesis” [Harris (1955)] when it states that the meaning of a word is defined by the nearby words, even without co-occurrence (and we hypothesize that semantics is linked to syntax in a way that depends on the specific field). The strength of the Harrisian hypothesis is attested to by the success of Latent Semantic Analysis (or Indexing) [Deerwester et al. (1990)]. Our experience, however, tends to show that local collocation is also very important.

6.3. Terms and collocations

In the present state of ACT, we still use the subject-verb and verb-object relationships. To that effect, we modified Carnegie Mellon University’s Link Grammar syntactic analyzer
[Sleator and Temperley (1991)] so it could accept our tagging. This obviously reduces its analytical power, but it also reduces its computation time enough to make it usable on large corpora.

Examples of the relationships we deal with, and their associated concepts, coming from the “DM” corpus are:

(accelerate:Verb,discovery:Object) \hspace{1cm} Algorithm
(detect:Verb,unusual-pattern:Object) \hspace{1cm} Nature of output
(agrawal:Noun,propose:Verb) \hspace{1cm} Algorithm-of-others (with a personal name)
(activity:Noun,monitoring:Noun) \hspace{1cm} Application
(actual:Adjective,data:Noun) \hspace{1cm} Input
(address:Verb,in:Preposition,section:Object) \hspace{1cm} My-paper-organization

6.4. ACT as a friendly interface helping the expert

The concept recognition step is entirely in the hands of the domain expert. It is important to stress that the entire process is strictly domain-specific. The expert has to define what the interesting concepts are. For example, in our biology corpus, this step is only comprehensible to a biologist, preferably one, who has a good knowledge of yeast molecular biology. For the introductions to DM papers, the specialist is the author of this chapter. Thus, the responsibility for defining a concept is assigned to the domain expert. Using the principles developed by Arnauld and Nicole\(^1\) in 1662, we can say that this expert certainly has a personal “definition in intention” of the concepts (possibly an unconscious one). When dealing with texts, we ask the expert to provide a “definition in extension” of how to describe the concept in a text.

Most these concepts are “commonsense” concepts, such as [ALGORITHM], meaning that the author describes an algorithm. Some were introduced while examining the texts, e.g., [BOASTING] (“self-congratulation”) and [AMISS] (where the author expresses doubts about the efficiency of algorithms). The specialist must browse the set of collocations in order to assign a concept to certain ones. For example, the sentence containing the words modest number is an instance of actual self-congratulation by an English-speaking author, hence its class, we called ‘boasting.’ ACT provides many clues for the specialist to help in the task of concept determination by means of syntactic forms.

Concepts already found in these sentences appear in color on the screen for the syntactic relations, and in underlined capitals for the terms. The list of concepts already found that contain the word data can be seen in the rightmost column [of the table at

\(^1\) The exact quote is: “J’appelle compréhension de l’idée les attributs qu’elle enferme en soi et qu’on ne peut leur ôter sans la détruire … J’appelle étendue de l’idée, les sujets à qui cette idée convient” [Arnauld and Nicole 1662/1965]. (I call understanding of the idea, the attributes it comprises in itself, and that cannot be removed without destroying it … I call extent of the idea, the topics to which this idea is suited [our translation].) It is remarkable that these notions are still used in data analysis.
the end of section 6.3]. If the user wants to add the expression *data in form* to the concept *NatofInput*, he or she simply clicks on *add to the taxonomy*. We evaluate that it requires up to 1000 h of intense work to manually categorize the concepts for the introductions to the English-language papers on DM. Clearly, creating a categorization *ex nihilo* is very difficult. ACT is especially useful for building a new categorization from another, already existing one in a similar domain. For example, we created a special categorization for the French-language texts in about 100 h, in spite of the very large differences between the two categorizations. Although we found large differences between the categorizations of concepts in the two corpora, we will not explore them here.

6.5. **ACT as an inductive program**

The last step in this process of concept characterization, and the one without which the whole process would be unbearably lengthy, is the inductive phase during which an existing categorization is automatically completed. At present, about 80% of the categorization relative to the introductions to the English-language papers on DM has been done by hand, and 20% by automatic induction [Kodratoff (2004)].

Concept recognition is a very large task because of the huge number of collocations that are actually used by domain specialists to describe a concept. For instance, around $10^5$ collocations were detected in the “human resources” corpus.

This task is therefore intractable if the expert does not own the necessary tools for obtaining these groups at the cost of a “moderate” effort, let us say some 100 working hours, possibly intense, put in by a top-level domain expert. Once this initial effort has been made for a specific domain, the induction tools that we have built, as described in this chapter, require only a moderate effort to shift from a field to one neighboring one, when applied by an expert in the field.

Our working hypothesis is that syntax is not the main tool for recognizing semantics. We rely rather on the more general and fuzzier notion of lexical dependency, as indicated by collocations. Thus, the domain expert must describe the concepts that are going to be used in the process of TM, by providing a “reasonable” amount of lexical dependencies enabling one to recognize these concepts within a text. Some user-friendly tools can help with this work, which, however, remains under the domain expert’s complete control. We say that these dependencies are defined in extension because we never ask our expert to provide an abstract definition of the concept; we request only a set of dependencies that the expert associates with a concept. The expert gives this set of dependencies a name. When by chance the concept name coincides with a noun that is a linguistic instance of the concept, we ask the expert to invent a new name, different from the existing linguistic instances, and we list these linguistic instances together with the concept name.

Induction works according to two modes: adding new lexical dependencies to a group defined as illustrating a concept and, conversely, eliminating dependencies that do not make sense to the expert. Lexical dependencies existing in the corpus under study can be grouped around a common pivot, associated with several different terms.
This constitutes a group of lexical dependencies. For example, all terms in the text associated with the adjective active constitute such a group. Depending on their size, the total number of these groups varies between $10^4$ and $10^6$. Note that in some linguistic theories, when the pivot is a verb, such a group of lexical dependencies is an instance of what is called the verb’s subcategorization scheme. Such schemes do not exist for adjectives, for example, in everyday language. On the other hand, users of genre specific-language may very well decide that some adjectives are semantically strong and that they impose their semantics on nouns with which they are associated. This notion is known as collocation, as already discussed.

To summarize, our sole goal is to introduce an inductive method that lets us assign new collocations to a given group of collocations that has been associated with a concept by the domain expert.

The first inductive work is the constitution of an exhaustive list of such groups of lexical dependencies containing at least $n$ lexical dependencies (that is to say of size $n$) existing in a corpus containing $N$ different words. We constructed an algorithm that builds these groups in $N^2$ time. This means an exhaustive search is always possible in a few hours of calculation time for $n \geq 3$ on a computer working at least at 1 GHz. Corpora with a very large and complex vocabulary, such as biology, may need to have their vocabulary reduced by lemmatization before this is attempted.

The following example presents two groups of lexical dependencies showing all the nouns gathering around a pivot, the adjective participative, and the nouns that come before the verb define (actually, the parser recognizes them as subjects of the verb).

- (participative:Adjective,atmosphere:Noun) Class-4731
- (participative:Adjective,mood:Noun) Class-4731
- (participative:Adjective,spirit:Noun) Class-4731
- (participative:Adjective,group:Noun) Class-4731
- (each-one:Subject,define:Verb) Class-3907
- (domain:Subject,define:Verb) Class-3907
- (leader:Subject,define:Verb) Class-3907
- (superior:Subject,define:Verb) Class-3907

We have observed that induction cannot lead to expert-acceptable results if it works with one only optimization measure. Generally speaking, each relevant point of view requests at least one measure. For instance, most of the statistically based inductive systems use two measures in the supervised case. One is used to optimize the search in the space of possible hypotheses, most often a criterion based on the minimization of variance, as in the k-means method or regression trees. The other measure is used to validate the model induced, and precision is the most often used for this purpose. In the unsupervised case, precision is not known before the induction; thus, one measure is used, generally the minimization of variance. In the case of language, we need to have at least three points of view; this is why, as a first step, we introduced three measures. Measures $m_1$ and $m_2$ are used to include new dependencies as instances of a concept,
and measure $m_3$ is used to withdraw a dependency as an instance of a given concept and to propose allocating it to another concept.

Measure $m_1$ measures a degree of cover between each group of collocations and each concept defined by the expert.

In the following example (no induction), group 1636 is completely included in the class defined as [COMMUNICATION] by the domain expert.

$$(\text{express:Verb,without:Preposition,fear:Noun}) \text{ communication}$$
$$(\text{express:Verb,without:Preposition,danger:Noun}) \text{ communication}$$
$$(\text{express:Verb,without:Preposition,reserve:Noun}) \text{ communication}$$
$$(\text{express:Verb,without:Preposition,constraint:Noun}) \text{ communication}$$

Here is an example with one possible induction. Two of the dependencies in group 2229 are included in the class [IMPLICATION] defined by the domain expert. One dependency in this group does not belong to [IMPLICATION]. Measure $m_1$ evaluates to what extent [IMPLICATION] covers 2229.

$$(\text{element:Noun,of:Preposition,motivation:Noun}) \text{ implication}$$
$$(\text{increase:Noun,of:Preposition,motivation:Noun}) \text{ implication}$$
$$(\text{part:Noun,of:Preposition,motivation:Noun}) 2229$$
$$(\text{increase:Noun,of:Preposition,motivation:Noun}) 2229$$

Measure $m_2$ measures the frequency of co-occurrence between a noun and the other nouns observed in the same group. In the example above, measure $m_2$ evaluates how often the noun part co-occurs in other groups with element and increase.

If these two measures are high enough, the collocation ‘part:Noun,of:Preposition,motivation:Noun’ will be added as a possible linguistic instance of the concept [IMPLICATION].

These measures give shape to the intuition that, if a computed group of collocations contains many dependencies identical to those of an existing concept ($m_1$ is high – this reflects the domain expert’s point of view), and if the noun, or non-pivot term, associated with this dependency is associated often enough with the other nouns in the group ($m_2$ is high – this reflects a property of the collocations existing in the corpus), then one can induce without consulting the expert that this dependency also belongs to this concept. The way these measures are computed, as explained below, is obviously temporary and will have to be refined as the number of experiences increases.

To calculate the $m_1$ of a group $G_i$, one must consider that, in any group of lexical dependencies $\{G_i\}$ a particular collocation can have three functions. It may belong to a concept $C_1$ and bring $G_i$ closer to $C_1$; or it may not belong to any concept and induction
may add it to \( C_1 \); or it may belong to another concept than \( C_1 \), thereby moving \( G_1 \) away from \( C_1 \). Let us call \( t \) the number of dependencies that a given group contains, let us call \( t_+ \) the number of dependencies in common with \( C_1 \), and \( t_- \) the number of dependencies in common with concepts other than \( C_1 \). For this group, we shall define its cover with \( C_1 \), by 

\[
m_1 = 100 \frac{t_+ - t_-}{t}.
\]

The coefficient 100 is only there to give \( m_1 \) the nature of a percentage.

In order to compute \( m_2 \) for a noun, noun \( i \), in a group \( C_1 \), consider the other nouns, noun \( k \), \( k \neq i \), found in this group. The set of the groups \( G_1 \) is browsed in order to count how many times noun \( i \) appears in the same group as noun \( k \). For each noun \( k \), count the number of times noun \( k \) and noun \( i \) belong to the same group; this is the value of \( m_2 \).

This way of measuring \( m_2 \) is obviously crude, but we shall keep it at this first stage. In particular, using LSA [Deerwester et al. (1990)] could provide interesting indications on the proximity of two words in the corpus. However, in order to avoid miscalculations due to polysemy, we must differentiate the meanings of polysemic words. This is precisely what the present step is for. We therefore plan to use LSA in a final refinement phase, and not during the initial phase described here.

The induction takes place, then, as follows: let us consider a given group of lexical dependencies, such as \( m_1 > 0 \), having at least one lexical dependency \( d_1 \) in common with concept \( C_1 \). Let us suppose that it also has at least one lexical dependency \( d_2 \) belonging to no expert-defined concept. The \( m_2 \) value of the non-pivot noun found in \( d_2 \) is then computed. Let \( mm_1 \) and \( mm_2 \) be the mean values of \( m_1 \) and \( m_2 \) for all groups and all nouns, and \( \sigma_1 \) and \( \sigma_2 \) the mean deviation of their distributions. One then adds \( d_2 \) to \( C_1 \) if 

\[
[m_1 > mm_1 + \sigma_1 \text{ AND } m_2 > mm_2] \text{ OR } [m_2 > mm_2 + \sigma_2 \text{ AND } m_1 > mm_1].
\]

This induction is performed without consulting the expert. This is a very positive feature for the expert, who need not work at this point, but it can lead to disaster if many dependencies are added that the expert later considers to be false. To avoid this drift, we shall spot any dependencies that contradict the existing definitions of the concepts and we shall point them out to the expert.

Measure \( m_3 \) measures how much an instance should be rejected or shifted from one concept to another. The intuitive meaning of measure \( m_3 \) is that, when a replacement is suggested with a positive value and the instance to be moved must also be deleted, then the expert is consulted to decide which concept this instance should belong to.

This measure is necessary because we introduce iterative induction steps, and we therefore need a way to keep the induction in check. This third measure is defined for this purpose, that is, in order to reject an instance from an already defined concept, defined either by the expert, or during previous induction steps. Let us measure \( m_{3+} \), the sum of \( m_1 > 0 \) such that dependency \( d_1 \) suggests adding \( d_1 \) to \( C_1 \) although it is already in \( C_2 \). This is the “strength of attraction” of \( C_1 \) for \( d_1 \). The value \( m_{3-} \), the sum of \( m_1 < 0 \) such that dependency \( d_1 \) recommends rejecting \( d_1 \) from \( C_2 \) is the “strength of repulsion” of \( C_2 \) for \( d_1 \). If \( m_{3+} \) is high enough and \( m_{3-} \) small enough, then \( d_1 \) is proposed to the expert, who must decide whether it should be removed from \( C_2 \) and added to \( C_1 \). The measure of the interest of this move, \( m_3 \), is a combination of the difference \( k_3 = m_{3+} - |m_{3-}| \) and the value \( m_2 \). Thus \( m_3 = f(k_3, m_2) \) and, in the present implementation, \( f \) is simply +, but this
function will require more work in the future. The only negative point of this measure is that the expert may be overloaded with questions. This is why $k_3$ is set to a weak value (say 10) at the start; if the expert judges it useful, $k_3$ is gradually increased.

6.6. Validation

This approach is quite new and, since we are still developing it, it has yet to be fully validated. We have however performed two kinds of validations.

6.6.1. Validation of the concepts

I built the “DM” corpus by collecting introductions to papers written in English by authors whose first language was English or French. After tagging the concepts found in these introductions, we used several DM techniques [Lerman and Azé (2003), Azé and Kodratoff (2004)] to exhibit structures relating the occurrences of the concepts, in the form of rules: if concept#i then concept#j. These rules are quite reasonable, and adequately describe what constitutes a good paper – or at least a good introduction.

6.6.2. Validation of the induction

The only corpus we have studied almost exhaustively is the “human resources” one. By examination, we found some 10,000 instances of the concepts deemed interesting by the psychologist. We applied extensional induction to this set of 10,000 instances, and noticed that many instances still needed to be added. More importantly, the system identified many existing concept instances as being contradictory because of their $m_3$ value. A detailed analysis of such “redirected” instances constantly demonstrated that the human expert had been wrong to attribute the given instance to the disputed concept.

We also ran the same experiment after having randomly eliminated 9/10 of the instances; of course, induction produced many more instances, but the same behavior described above was observed again.

7. Conclusion

Almost all the Internet advertisements for companies doing computational linguistics claim to be able to find or use concepts, showing the industrial importance of solving the problem of concept detection. This chapter presents a technique (and the associated problems) enabling experts to build their own concepts, and enrich them using the texts produced by their field.

Thus, we propose a partial solution to the old problem of deriving semantics from syntax, with this restriction: we actually reveal the particular semantics associated with the syntax of a specific technical field. In spite of the strong cognitive appearance of this problem, the reader can see that we never tried to really “understand” what a
concept was: for us, a concept is a set of linguistic instances, a few of which were provided by a human.

Being able to recognize concepts in texts amounts to simulating some deep understanding of the text. This does not directly solve any other problem, but it obviously provides a kind of knowledge that can be integrated into other approaches. Let us give a few examples of such uses in the domain of coreferences in a technical text.

A characteristic feature of technical texts is that almost nowhere is a physical measurement between two objects or events given with the specific names of these objects. Our experience is as follows: the author introduces the objects or events, describes them at some length, and explains an experimental setting, describing how the measurements were performed on these objects or events. Finally, the author uses a hypernym of the objects or events, such as “this type of object,” and describes the measurements that were performed. This classical coreference problem appears to be very difficult for two reasons. Many sentences (those describing the experimental setting) may separate referent and referred, and the referent word has usually never been used before since “this type of object” is described by its actual name. For instance, in molecular biology, an author may speak of the “sequences flanking the core 8 bp TATA box” and subsequently refer to this concept by “these regions.” That the “sequences flanking…” are regions is taught in elementary-level biology courses and should be made explicit in building the ontology of concepts for this field.

This explains why we ask the expert to provide a concept name, and a list of the possible nouns used by domain experts to designate this concept. Note that these nouns are hypernyms of the concept’s linguistic instances. We recommend to preventing the expert-given name of the concept from coinciding with any of these nouns because such nouns are yet another set of linguistic instances (at a higher level of generality) while the name of the concept is a kind of metaknowledge used to structure the knowledge.

The linguistic instances, i.e., the collocations we obtain from the expert, contain an instance of the objects or events described by the author of the paper, either directly, or by inclusion. Our approach based on concept recognition enables us to recognize both objects and events, and to find the coreferential hyponyms. We believe that our approach is the only one that can succeed in this task, which explains why information extraction for filling up databases has generally been considered to be still impossible.

Acknowledgments

The lexicon was built by Thomas Heitz; the tagging language was written by Jérôme Azé; the tagging interface (ETIQ) is due to Ahmed Amrani; the work on terminology/collocations was done by Mathieu Roche, while Thomas Heitz wrote the corresponding interface (EXIT); the ACT software is due to Jérôme Maloberti and Ahmed Amrani. The software for extensional induction was written by Jérôme Azé.
References


Harris, Z. (1955), From phoneme to morpheme. Language 31:190–222.


Chapter 43

CLASSIFICATION AND CATEGORIZATION IN
COMPUTER-ASSISTED READING AND TEXT ANALYSIS

JEAN GUY MEUNIER AND DOMINIC FOREST

Université du Québec à Montréal

ISMAIL BISKRI

Université du Québec à Trois-Rivières

Contents

Abstract 956

1. Introduction 957

1.1. CARAT: General presentation 957

1.2. Difficulties with the technology 958

1.3. The nature of reading and analyzing a text 959

2. Definitions of classification and categorization for CARAT 961

3. Text classification and categorization 962

3.1. Text classification 962

3.2. Text categorization 962

3.3. Computer text classification and categorization 963

4. Methodology for text classifying and categorizing 963

4.1. Steps 1, 2, and 3: From a text to a matrix 964

4.1.1. Step 1: Identification of units of information and domains of information 964

4.1.2. Step 2: Cleaning and filtering 965

4.1.3. Step 3: The matrix 966

4.2. Steps 4 and 5 967

4.2.1. The classification process 967

4.2.2. The categorization process 968

4.3. Step 6: Navigation 968

4.4. Step 7: Evaluation 968

5. Applications in CARAT 969

5.1. Thematic analysis 969

5.2. Categorical exploration of philosophical texts 970

5.3. Content analysis 972
Abstract

Automated computer classification and categorization have been successfully applied in the field of document retrieval and text mining. These methods are increasingly used in the field of computer-assisted reading and analysis of texts (CARAT) in the field of the humanities and social sciences. Here, the primary aim is not to tag the document for better recall, but to help a specialized reader and analyst to navigate through the themes, concepts, and content of a textual corpus according to a particular research objective. This chapter presents some of the methods and applications of these techniques.
1. Introduction

1.1. CARAT: General presentation

In the early 1960s, the computer appeared to be a revolutionary tool for computing mathematical symbols – even numerical ones!

Scientists in AI saw computers as machines that manipulated symbols. The great thing was, they said, that every thing could be encoded into symbols, even numbers. *Newell (1983, p. 196)*

Surprisingly though, the real impact of computer technology was not on the processing of numerical symbols but on work with a wide range of other types of symbols, such as natural language symbols.

Indeed, the computer has been widely used for text processing, so much so that today most of the processing done by computer is applied to text (Internet, e-mails, documents, etc). Since the early days, research in cognitive sciences and information technologies (IT) has had a major impact on the reading and analysis of texts. And the field of the humanities has been the beneficiary.


The encounter between these various disciplines and computer sciences and technologies has given rise to an original research field called Computer-Assisted Reading and Analysis of Text (CARAT) [Meunier (1997), Bernard (1999), Popping (2000), Hockey (2001)]. This research field differs from the artificial intelligence (AI) approach to discourse analysis [Hobbs (1990)] or automatic reading [Ram and Mooreman (1999)], where the objective is to have the computer simulate some type of “understanding” of a text in some specific application or process (inference, summary, knowledge extraction, question answering, e-mail routing, etc.). It is also different from information retrieval [Salton (1989), Salton and McGill (1983)] or hypertext technologies [Rada (1991)], where the objective is to find documents using a particular query (or to navigate through documents). In CARAT, literary critics, philologists, content analysts, philosophers, theologians, psychologists, historians, and many other types of professional text readers (lawyers, journalists, etc.) require computer tools that assist them in their own, often personal, reading and analysis. And they cannot accept automatic text “interpretation” tools in any form whatsoever.
Computer technology now offers more and more possibilities to assist the reading task. Archives of electronic texts are now common (Oxford Text Archive, Brown Corpus, Perseus, Gallica, Frantext, etc.). They are increasingly likely to be critically edited and standardized (SGML, HTML, XML, etc.) and can be explored with “intelligent” tools such as navigators, search engines, hypertexts, etc. [Condron, Fraser and Sutherland (2001)].

Technology has also been very constructive in assisting with the analysis process. In fact, one can distinguish different generations of this technology. The first one (1950–1970) opened the era of electronic text capture. A second generation (1970–1980) offered tools for the description and manipulation (extraction, concordance, statistics, lemmatization, etc.) of these electronic texts. The third generation (1980–1995) started the standardized tagging (SGML, TEI initiative) and linguistic processing (syntax, semantics, discourse and rhetoric, etc.) of the text. A fourth one (since 1995) introduced sophisticated mathematical models (classifiers, neural nets, categorizers, etc.) of the text. Today, a host of such tools exist: from the hybrid to the very specialized, from laboratory prototypes to industrial applications (Concord, Word Cruncher, TACT, Cobuilt, Word Smith Tools, Lexa, Cosmas, Tropes, Intext, Alceste, Sphinx, Hyperpro, Atlas/TI, NU*DIST, etc.). CARAT has even received a buzz name in the commercial — “Text Mining” — for which a multitude of commercial tools exist.

1.2. Difficulties with the technology

Unfortunately, the traditional communities in the humanities and social sciences do not recognize the full importance of this technology, except perhaps for critical editions of texts, qualitative content analysis projects, or historical archiving. One can invoke many reasons to explain these difficulties. One is the poor ergonomics of these technologies, which makes them difficult for many users to learn and use, particularly in the humanities and social sciences [Unsworth (2000), p. 2]. Most of the available technologies have been developed by communities external to CARAT (linguistics, artificial intelligence, data processing, information retrieval, etc.) and for specific objectives. A second reason relates to the limited sets of tools available for text analysis; often a researcher is confined to text transcription, text encoding, lexical analysis (concordances, collocation), linguistic analysis (lemmatization, tokenizer, morphological taggers), or statistical analysis. Third, many commercial tools are often geared more to information retrieval than real assistance for reading and analysis of text. Finally, the technology is often closed. That is, it is proprietary, rigid, and nonmodular and, as is often emphasized, not well adapted to the dynamics of CARAT projects.

We believe, though, that there are more profound theoretical reasons for these difficulties. There is a lack of understanding of the real role played by this technology. The mainstream community still has the wrong image: that the computer is doing some type of text interpretation. It has difficulty in seeing the computer as being just an assistant in these processes. There is also a misunderstanding of the linguistic aspects of a text, its reading, and its analysis. Rastier (2001) reminds us how much a text as a whole differs from a sentence in a language; consequently, it must be approached with tools other than
grammar and logic [Mann and Thomson (1988)]. Finally, the classical CARAT tech-
nologies, except for the ones related to standards and protocols [Sinclair (2002)] have not
really integrated the modeling tools of the software engineering languages developed in
the last 15 years [Rumbaugh et al.(1991), Lévesque (1998)], which allow for modularity,
reuse, exchanges, collaboration, and even automatic code generation. If a renewal of
the use of this technology is to happen, we must have a more rigorous idea of what read-
ing and analyzing a “text” with the help of a computer involves. Therefore, a more the-
oretical and formal foundation for CARAT technology has to be explored.

1.3. The nature of reading and analyzing a text

Any such renewal must be based on a better understanding of the real nature of a text.
The term “text” is very ambiguous. In certain contexts, it may mean a paper object (e.g.,
a Gutenberg Bible). In others, it may mean the abstract contents (e.g., the Bible as a
text). Sometimes it is a linguistic object (as in, writing a text is different from writing a
sentence) or a psychological object (as in, the understanding of a text).

A text is in fact a complex semiotic object. Even if a text presents itself as a sequence
of natural language symbols, it is more than a “linguistic” object in the sense of more
recent linguistic theories. It is not a well-formed output of a formal grammar. Also, it is
not identical to discourse. Indeed, the same text can be read, pronounced, or enunciated
in many different contexts (cf. the Bill of Rights), and each time it produces a different
discourse effect.

A text contains different dimensions such as argumentation, narration, description,
demonstration, dialog, theme, etc., none of which are strictly grammatical phenomena
in themselves, although they may present some strong regularities that belong to their
genre [Rastier (2001)], their rhetorical structure [Mann and Thomson (1988)], or their
logical structure [Hobbs (1990)]. In fact, if a text was a language in the strict grammat-
ic(algorithmic) sense, then, in learning to produce and recognize sentences and struc-
tures of a particular language, one would also learn to produce and recognize textual
structures. For instance, if one knew how to produce and recognize Danish sentences,
one would also know how to produce and recognize Danish legends!

This means that reading and analyzing a text is not strictly an algorithmic process
[Eco (1965, 1992)]. We would add that it is probably not even a computational one
[Meunier (1997)]. Reading or analyzing a text is more akin to complex procedures of
heuristics and pattern recognition strategies than to pure grammatical and algorithmic
procedures. This access to the textual content is, in fact, a complex interpretive process
rooted in various dimensions, such as linguistic and real-world semantic structure,
infrence, pragmatic memory, culture, social interaction, knowledge repositories, etc.
Hence, it is in the nature of a text not to reveal itself in its totality. Every text is poly-
valent. Its meaning varies according to the context of reception [Jauss (1978)] or the
reader [Eco (1965)], and it does not necessarily reveal itself in a first reading or even in
a first analysis (cf. the Bible or the Koran). The hermeneutic tradition has emphasized
this point for at least the last century [Schleiermacher (1808–1809/1987), Heidegger
(1962), Habermas (1984, 1987), Gadamer (1996)]. And contemporary theories of reading expand on this thesis [Barthes (1970), Fish (1980), Iser (1985), Gervais (1993)]. In light of these points of view, if we wish to use a computer to assist with the reading and analysis of text, we must try to “follow” as closely as possible the strategies used by human readers and analysts. And this is still more the case if these reader/analysts are experts such as those in the humanities and social sciences.

There are many strategies for a good, deep reading of a text. Finding them is the aim of many psychological, philological, and computational theories of reading. Among these many strategies, two major ones are often used: a classification process and categorization process. Indeed, a reader or an analyst does not simply follow the natural course of a text, line by line, chapter by chapter. He aims to find some underlying, implicit pattern or structure of some sort that is called, in different disciplines, a theme, a concept, an argument, etc. Text analysis is, in one sense, a second-order reading that classifies or categorizes textual objects.

In the classification process, the reader, through various clustering, regrouping, or reorganizing techniques, tries to discover classes of relevant textual objects on the grounds of some type of similarity criterion. For instance, in literature, one may want to group together all the relevant passages of Shakespeare on a particular theme ([JEALOUSY], for example). The various text passages may or may not express this theme explicitly. It may start to emerge at the beginning of the corpus, and only be well exposed and expressed at the end of it. It is here that a classification procedure becomes a highly relevant tool for analysis. It is, in fact, an implicit practice in many lexical, semantic, narrative, or stylistic analyses.

In performing these classification and categorization procedures, an expert reader will call upon many heuristics that are part of his expertise, cultural background, erudition, objectives, etc. And each reader and each expert has his or her own repertoire of techniques, procedures, and best practices, which underlie his or her personal and subjective interpretation. In fact, they belong to his signature as an expert. And it is obvious that this highly complex interpretive process cannot be translated into easy and clear-cut algorithms. And it cannot be easily applied and used by a computer so that the reading and analysis becomes automatic! But it does not follow from this that the computer cannot be useful in assisting these processes. Indeed, it can play a productive role, but only if it is faithful to this interpretive process and well situated in the cognitive activity of the reading and analysis of a text. CARAT appears relevant from this perspective. It has in
fact become an important tool in various social sciences and humanities disciplines that rely on rich reading and analysis of text. In this context, CARAT offers a set of algorithmic tools assisting the construction of interpretive paths in reading and analyzing texts. And scholars in these fields will accept this computer technology if it really acts as an assistant in their own intellectual examination of a text and does not try to substitute itself for them.

2. Definitions of classification and categorization for CARAT

If CARAT is to be an assistant in the process of reading and analyzing texts, one must better define the concepts of classification and categorization. Is classification categorization and vice versa? Unfortunately, it is not easy to define these two concepts, which have received numerous definitions in different scientific disciplines. Let us distinguish some of the most important differences between these two concepts because, in the scientific literature, the concepts of class and category are often amalgamated. A category is a class and vice versa. But no matter what names are used, one must distinguish the underlying concepts. A first concept sees a class or a category defined as the result of some formal operation: the partition of a set of objects. This operation must offer the conditions under which a set of such objects are related among themselves in order to become a class. This is usually the definition used by mathematicians and logicians.

The second concept sees a class or a category as a symbolic label, that is, the name given to the class produced by some classifying process. This is the concept used in linguistics, information science, and often in computer science. Here a category is often a labeled class.

The third concept defines a class or a category as a cognitive state – it pertains to the agent that cognitively performs the classification and categorization processes. This is the focus of interest for much psychological, anthropological, and computer sciences research.

The fourth concept understands a class or a category as a morphism, that is, as a type (in the hierarchy) of structure that classes may entertain among themselves. This is the focus of structural (geometrical) and mathematical modeling of classes.

These conceptual distinctions lead to various types of scientific inquiries and, even though they are often related among themselves, they must not be confused. The problem of differentiating the varieties of [ROSES] is a different problem from that of naming them. And this is not identical to the problem of situating [ROSES] among the various flower species (taxonomy). This is also different from the psychological cognitive process by which a nonbiologist or an experienced gardener can identify that flower. Even though we can always use the terms classes and categories, the important thing in each case is to see whether they are applied to the same or different operations, and hence, pertain to identical or different inquiries. And the solution to one problem cannot be a solution to the other.
3. Text classification and categorization

Given these various possible distinctions between class and category, let us now see what takes place with CARAT. Here, we can distinguish two different research programs that have been the subject of attention in the past decade.

3.1. Text classification

The first one – text classification – pertains mainly to the first definition of classification, that is, classification as clusters, sets, etc., applied to objects – in this case, to textual entities (sentences, documents, books, Internet sites, etc.). “Text classification is the automated grouping of textual or partially textual entities” [Lewis and Gale (1994)].

In more technical terms, text classification is defined as an operation that is applied to textual entities by which equivalent classes are built. Classification is therefore a process by which textual information is clustered together according to some criteria. A query on the Internet is a simple, basic example of this. A query recalls all the textual entities (sites) that contain the specific words in the query.

But the more interesting classification techniques are procedures where the criterion is slightly more complex than that. For instance, in a large text corpus (the Shakespeare corpus, for example), one may want to group together all the relevant textual entities that talk about different themes.

The research objectives in text classification are mainly to find techniques that build these classes either in a top-down manner (by a list of predefined conditions) or in a bottom-up manner (by discovering inductively the conditions under which the textual entities are similar). Such classification systems deliver sets of textual entities. They never produce labels that could either name or label the classes themselves.

3.2. Text categorization

The second research program relates to categorization. It pertains more to the second definition of category, although, implicitly, it also relates to the third and fourth definitions. As Sebastiani (2002; p. 1) puts it, text categorization is “the activity of labeling natural language texts with thematic categories from a pre-defined set.” In other words, text categorization is “the process of assigning entries from a predefined vocabulary” [Ruiz and Srinivasan (1998), p. 59].

Here, the aim is to add labels or symbolic tags to the classes that are constructed or found. Usually, as defined above, the set of tags is predefined and belongs to some system of classification. This is a top-down approach. But in certain cases, the labels themselves are not predefined and must be discovered by some means. This is more typical of ontological and taxonomic applications.

Hence, text categorization is more than text classification. Categorization relies on a classification process, but adds labels to the classes, that is, it attributes to them some
type of predefined tags. These tags describe some aspect of the “semantic” content of a class of textual entities. It is in the choice of the labels that the cognitive and structural dimensions of categorization are called upon.

For instance, all similar segments of text classed together could be tagged by category terms such as [LOVE], [HATE], [JEALOUSY], etc. They could also be tagged as [PASSION], [EMOTION], or [ACTION]. And in so doing, one sees that there is an implicit cognitive and structural dimension at work. There is a difference between tagging a set of textual entities as [HATE] or [LOVE] versus [PASSION] or [ACTION].

3.3. Computer text classification and categorization

Both of these research strategies can be performed manually; in fact, they have been for centuries. And it is still the most popular procedure today. Librarians still classify and categorize books manually. Most of the thematic analysis in scholarly literature or philosophy is done in this way. Even qualitative content analysis of interviews, testimonies, or historical documents follows the traditional procedures of reading and analysis.

But more and more, these classification and categorization techniques are being explored in a computational framework. Indeed, some research programs aim to find and evaluate various algorithms that can classify and categorize textual entities by means of computation. This research program has become quite a stable paradigm in itself, and activity is quite lively in the fields of information and knowledge management technologies (often called “text mining technologies”), for which a rich and relevant literature exists [Lewis and Gale (1994), Kodratoff (1999), Sebastiani (2002)].

Our own research program is to see whether these algorithms can be successfully applied in the context of the reading and analysis of texts. In other words, we are trying to find out whether they are useful in assisting reading and analysis practices in the humanities and social sciences. We intend here to explain how these automatic classification and categorization approaches can be heuristically applied to CARAT, as this is not immediately obvious. In this field of application, these techniques cannot be an end in themselves. No text reader or analyst would find it very useful to subject her corpus to classification and categorization just for the sake of it. But we believe that these techniques can be useful if they are linked to the dynamic process of text interpretation.

In the field of the humanities and social sciences, these techniques are used, mainly in European academic circles, for statistical discourse analysis. But they have not really taken root yet in the American tradition of computer text analysis.

4. Methodology for text classifying and categorizing

Text classification and categorization rely on a similar methodology, at least in their first steps, but they differ in the later steps. Here, we present this technique in seven main steps.
4.1. Steps 1, 2, and 3: From a text to a matrix

4.1.1. Step 1: Identification of units of information and domains of information

The first step consists in defining what the textual entities or textual input for the classifying and categorizing algorithms will be. We distinguish between two different types of textual entities. Each of them will be required later on in the two processes. The first type of textual entity is what we call units of information (UNIFS). The second type is domains of information (DOMIFS).

4.1.1.1. Units of information (UNIFS). The first type of basic unit of information in a text is linguistically grounded [Kucera and Francis (1967)] and is usually defined in terms of “words” or some of its linguistic variants. For instance, we can find:

- **Basic words** (any sequence of letters separated by blank spaces)
  “John loves Mary”: John, loves, Mary
- **Lemmas or stems** (words reduced to their basic forms)
  loved _ love (or lov)
  intermediately, intermediate, intermediates _ intermediate
- **Morphemes** (words that have been distinguished according to their syntactic categories)
  cry (noun) vs. cry (verb)

One very useful type of unit of information is composed of compound words (phrases) that are grammatically grounded. These phrases can be made of nouns (such as, jet set, table set, chess set, table napkin, chest pain) or verbs (such as, kick the bucket, step on, etc.). Others are purely based on sequences of words or collocations (word grams) [Choueka (1988)]. For instance, in the sentence John loves Mary today, one could define John, loves, loves Mary, or Mary today as units of information.

Another type of unit of information may be sensitive to syntactic or semantic information. For instance, the type word house can be distinguished as house_{noun} or house_{verb}.

A second basic type of unit of information has recently been used. These are called *n*-grams and are composed of *n* symbolic characters [Damashek (1989), Cavnar and Trenkle (1994)]. For instance, in the sentence John loves Mary a three-gram will take sequences of three letters (Joh, ohn, hn_, etc.). These sequences of *n*-grams can also be constrained by certain probabilities [such as a Bayesian approach: Church et al. (1989)]. That is, not all sequences of letters are retained; only those with a certain probabilistic characteristic.

As one can see, all these UNIFS must be automatically identified by some algorithm. This is what tokenizers and parsers (syntactic, morphological, or semantic) do. Naturally, each algorithm will encounter many complex problems. Whether they be words or *n*-grams, the choice of the right UNIFS depends on the aim of the operation.

Each choice of unit of information has its own advantages and drawbacks. If the UNIFS are all basic words in all their inflectional variety, the list of UNIFS is more respectful of the language but the list can easily expand into tens of thousands of words. Often, the choice of one or another kind of unit of information may depend
4.1.1.2. Domains of information (DOMIFS). The second type of information to be identified in this first step is the domains of information (DOMIFS) (the text fragments on which a classifier will be applied). These fragments may be phrases, sentences, paragraphs, pages, chapters, documents, or fixed sequences of words.

To concretely identify these DOMIFS in a corpus, some type of algorithm is needed. In continuous texts, this will be done by a segmenting program that operationally defines what is to be retained. In information retrieval, these DOMIFS will often be documents or Web pages. In CARAT, the DOMIFS are usually paragraphs.

What is the best fragmentation possible, and for what purpose (a page, a document, a title, a sentence, a line, etc.)? The answer depends on the aim of the classifying procedure. And here, only repeated experiments will determine what length or type of fragments fits best for a particular type of analysis. In our own experiments, we favor medium-length paragraphs (50–1000 words). This range seems to represent the average length in which an author has time to introduce and develop a theme.

Another problem is the algorithm or set of rules necessary for identifying the chosen fragments. The information retrieval (IR) tradition applies an easy strategy by taking the document as a basic fragment. An “abstract” is even easier to identify. Unfortunately, this is not true of ordinary plain text. Indeed, in the humanities, classical texts are standardized with XML tags. This helps delimit the text but makes classification and categorization more cumbersome.

Also, in these domains, the nature of a line, a paragraph, or a page is not clear; this is even truer when one considers a sentence or a phrase. Each choice brings its share of problems and computer processing costs.

4.1.2. Step 2: Cleaning and filtering

The next step consists of cleaning the corpus and filtering some of its information units (UNIFS or DOMIFS). The aim of this step is to produce a set of the most relevant information units.

For instance, in a classical edition of a literary book, one may not wish to retain the title page, the word “chapter”, etc., as basic units of information; these all constitute meta-information that a good edition will integrate.

Because of the mathematical structures of certain classifying techniques, it may not be relevant to keep very low-frequency words. It is also usual to remove functional words (or stop words) such as articles, prepositions, punctuation marks, etc., from the text.

Even though this step is not theoretically very complex, it is necessary, at least for the classification and categorization processes, because it considerably reduces the amount of data to be processed. But it will often be a big surprise for the novice at CARAT to find out how long and complex the cleaning and filtering phase takes. It will call for many decisions, not all of them easy. Moreover, in some cases, due to the
idiosyncratic nature of some of these decisions, the operations may have to be done manually.

4.1.3. Step 3: The matrix

In the third step, the textual units and fragments are transformed into a matrix using the Vector Space Model [Salton and McGill (1983), Manning and Schütze (1999)] (Figure 1). Here, each segment, with its information units, is seen as a vector where the informational content is translated as a property of this vector. The values given to each entry in the matrix represent the attribution of this property to the vector. The type of values given depends on the classifying model chosen (e.g., presence, absence, fuzziness, weighting, etc.). All functional and subjectively irrelevant words can also be eliminated from the text. This can be done either manually (according to a goal set) or automatically (using rules or a dictionary).

The matrix can represent a direct count of the UNIFS in each segment. But as the literature has shown, many variants are possible.

Thus the whole text is transformed into a vectorial space. This allows the use of all the mathematical tools that can be applied on this space. This is the strength of the classification and categorization approach. The text is parsed not linguistically but mathematically. From here on, classification and categorization techniques are different in their methodology, albeit slightly related.

![Fig. 1. The matrix fragments of text–units of information.](image-url)
4.2. Steps 4 and 5

4.2.1. The classification process

In the classification procedure, the fourth step consists of applying some mathematical or logical (rule-based) “regrouping techniques” to the set of vectors (the matrix). There are many such techniques. In the literature on mathematical text classifiers, many types of classifiers have been proposed and explored, all of which have their parameters, and hence, their fecundity and limits. One successful implementation of these types of models has been the SMART system [Salton (1989)] in the information retrieval community. The classical types include statistically oriented techniques: clusterizers, correlators, and factorial analysis [Reinert (1994)]; principal component analysis [Benzecri (1973)]; K-means [Hart (1968)]; neural network classifiers [Kohonen (1982, 1997), Carpenter et al. (1991)]; genetic algorithms [Holland (1975)]; and Markovian field classifiers [Bouchaffra and Meunier (1995)]. Some recent important variations in these models are to be found in Latent Semantic approaches [Deerwester et al. (1990)] and support vector machines [Cristianini and Shawe-Taylor (2000)]. Some more probabilistic techniques have also been tested, such as Bayesian classification or machine learning techniques [Kodratoff (1999), Sebastiani (2002)]. Finally, many recent research projects are exploring the potential of hybrid techniques. A good example of such techniques is found in Nauck (1999), who developed a neuro-fuzzy classifier which combines a traditional neural net with fuzzy logic concepts and techniques.

These techniques group together textual segments (domains of information) under some criterion of similarity. At the end of the classification process, the system offers the reader classes of similar segments based on similarity criteria. All of these mathematical techniques have been applied to textual information processing and discourse analysis. Although they have their limitations, the results have been good and they compare very positively with the more linguistically oriented techniques. Their great advantage is the savings in processing time. They have also been shown to be essential in processing large textual corpora.

Normally, classification is the last step. But in CARAT, one must further explore the classes found. So, in a fifth step, the classes produced are analyzed and some of their properties are extracted, depending on the objective pursued. The main objective of this fifth step is to find in the classes of segment a way to filter them out, summarize them, and present them transparently. In this process, one could focus, for example, on the semantic information present in a particular class of segment (the main topic, the semantic field, etc.).

This is done using logical and mathematical operations (i.e., statistics, set theoretical, decision or semantic rules, etc.) applied either to DOMIFS of the classes or to their lexicon. The means by which the content of the classes is read and analyzed is not a

---

1 Every 2 years, an international conference is held for researchers in the social sciences and humanities who apply statistical and mathematical techniques applied to computer text processing (Journées internationales d’Analyse statistique des Données Textuelles (JADT)).
well-defined procedure. It is an ongoing and important research problem. It is here that interpretive analysis is really at work; and it is from here that the true content analysis of the text really emerges.

4.2.2. *The categorization process*

In the categorization procedure, the fourth step consists in defining, from a set of predefined categories, the labels that will be used to “tag” the fragments of text (each vector of the matrix). This set of tags is taken as a working hypothesis for the expert reader and it can originate from various sources (thesaurus, classification systems, content organization, dictionaries, taxonomies, etc.). The system is then trained by simulation: the computer system follows an expert reader who manually tags a sample set of segments. From this tagging, the categorizing algorithm then “learns” what “count” as exemplars of a particular tag. There are a variety of algorithms to learn categorization in this way. Some are purely mathematical (perceptron, Rocchio’s algorithm, etc.); some are a mix of rules governing decision trees (cf. the Ripper algorithm); and some are typical machine learning strategies [Michalski (1983)] grounded in various external tools such as Word Net. In the fifth step, once the learning is consolidated, the results are projected onto the whole text. By itself, the system then tags the remaining segments of the text by categories. This is done through the matrix built in the second step. The categorization techniques are then applied to the matrix. That is to say, because the system “knows” which category a sample vector belongs to, it finds all segments that are similar to the ones it has learned and automatically projects the same label on them.

4.3. *Step 6: Navigation*

Once the corpus has been correctly classified or categorized, the system needs to offer mapping techniques [Barry (1998)] that present the organization of the classes of textual entities, and their lexicons or lists of categories [Spence and Press (2000), Fayyad, Grinstein and Wierse (2001)] in a user-friendly way. Although theoretically basic, these techniques offer real assistance with the interpretive activity of reading and analysis.

4.4. *Step 7: Evaluation*

Finally, in the seventh step (for both classification and categorization), evaluation measures (precision, recall, accuracy, etc.) are applied to the results. These evaluation techniques, implemented in algorithms, have been (and are still) applied to many information processing domains. They are used in information retrieval, knowledge engineering, knowledge extraction, natural language processing, ontology building, data and text mining, etc. For instance, in information retrieval techniques, relevance feedback, precision, and recall have become the norm. For some classifiers, comparative strategies have also been applied, such as Van Rijsbergen’s measures, which compare the elements of classes.
We believe that this type of methodology is difficult to apply in CARAT, for, despite appearances, reading and analysis of text are not objective processes but are highly subjective. The problem is related to the complexity of the interpretive activities of reading and analysis. As we have already mentioned above, because of the great variation in text interpretation, it is not easy to find standards that could serve as parameters for these evaluations.

For instance, what set of hyperlinks could be objective enough to be taken as a benchmark? Personal experience shows that one often navigates through websites in a haphazard fashion. This is not usually a result of the poverty of the technology but the effect of the discovery process a reader carries with her. For this purpose, we think that a better adapted evaluation process can be provided that will be more faithful to the idiosyncratic strategy a particular user follows, in her reading and analysis of text.

Often, the benchmarks used are the classical interpretations given to some important textual corpora. In the following section, we shall illustrate, from our own research, some heuristic applications in CARAT.

5. Applications in CARAT

We will now illustrate how the classification and categorization techniques presented above can be heuristically applied in the field of CARAT. We have chosen a subset of cases taken from our own past experiences. They show, we hope, how these computer techniques can assist in the process of reading and analyzing texts.

5.1. Thematic analysis

Our first application for CARAT illustrates how the classification approaches can be used in a thematic analysis of philosophical texts. Indeed, one of the main basic philosophical interpretation tasks is conceptual analysis, as developed under a specific theme. For instance, a philosopher may want to explore the concept of [TRUTH] in Descartes’s writings, and discover the various semantic fields in which it operates. From a computing point of view, such a thematic analysis task is more complex than information retrieval. The reader does not know before the analysis what the span of the lexicon will be in which this theme is expressed in the corpus. So the classification approach can be a very heuristic tool to support the discovery and exploration of a particular theme.

In this experiment, we applied a classification approach to a French philosophical text (Descartes’s *Discours de la méthode*). The original untouched text (except for basic printing noise) contained 21,436 words. The text was subjected to the various steps of the classification methodology described above. At the fifth step, the process produced classes of segments according to certain similarity criteria. In this experiment, we chose the ART1 neural network classifier. From the classes obtained, a lexicon of the segments was extracted. Figure 2 shows the lexicon for each thematic class.
The analysis starts by choosing a word that is present in many classes. This becomes the leading “thematic word.” In the following illustration, we have chosen the word (or concept) *connaissance* (‘knowledge’), which appears in many classes.

Figure 2 shows that, in fact, the word *connaissance* is used in various semantic fields. For instance, in one of the classes, the vocabulary surrounding the term *connaissance* relates to the metaphysical question in Descartes’s writings. In another class, it operates in more of a mathematical environment. It is here that the expert reader introduces his interpretive abilities. The computer has merely classified the text segments according to certain similarity conditions that are seen as expressing dimensions of a semantic context.

With this preliminary classification function, the reader can then “direct” the analysis according to the chosen theme. And from then on, he can explore other themes found in Descartes’s philosophy. For instance in the following example (Figure 3), he may switch to another concept inside a particular class [e.g., *pensée* (‘thought’)] and start a new thematic path. This strategy is then applied to the entire text. Other results [Forest and Meunier (2000), Forest (2002)] have shown that this strategy can be applied to many other concepts.

5.2. **Categorical exploration of philosophical texts**

Another application of the classification and categorization techniques in the humanities consists in finding in a corpus the various segments of a text that are representatives...
of certain conceptual categories. For instance, one might be interested in finding the various text segments that are relevant to the thematic category [TRUTH].

In this particular application, the expert reader has the system learn the passages of a text that she believes correspond to a conceptual category that she intuitively thinks underlies a particular passage. This is done on sample segments of the text. Afterwards, the learning is projected onto the whole corpus. This procedure is typical of the learning phase of classical categorization techniques.

This strategy is used because, in many highly technical and abstract texts, one does not and cannot know beforehand how an author will explore a particular theme. And the
analyst would like to explore whether some of her intuitions about the theme are really at work in the text. For instance, what themes are the most important ones in the corpus? How is a particular theme realized in the corpus? Are there other subthemes related to a chosen one? Because of the complexity of the information processing involved, the computer becomes a helpful heuristic tool in exploring certain intuitive categorical themes in the text.

We applied this strategy to Bertrand Russell’s text *The problems of philosophy* [De Pasquale and Meunier (2003)]. The application consisted in having the system discover, through learning, passages corresponding to specific categories (as opposed to terms) such as [KNOWLEDGE], [ETHICS], [SPIRIT], even though the passages did not actually contain these words.

After the learning process (carried out on samples) was completed, the system, through a simple perceptron classifier, managed to find many sentences expressing the chosen categorical themes. These segments were then presented to the expert for analysis and evaluation. For example, the following segment was found to be relevant for the category [KNOWLEDGE]: “In this respect our theory of belief must differ from our theory of acquaintance, since in the case of acquaintance it was not necessary to take account of any opposite. (2) It seems fairly evident that if there were no beliefs there could be….” But the system also rejected the following segment: “Some relations demand three terms, some four, and so on. Take, for instance, the relation ‘between.’ So long as only two terms come in, the relation ‘between’ is impossible: three terms are the smallest number that renders it possible. York is between London….”

The matching of the sentences and the categories relies not only on the presence of certain words but on the simultaneous presence of certain related terms such as acquaintance, knowledge, truth, reason, etc.

This type of very simple dynamic categorical learning may not be suitable for information retrieval and indexation but is undeniably valuable in the heuristic exploration of a thematic category in a humanities text.

5.3. Content analysis

In social sciences, an important computer application related to CARAT is qualitative content analysis of discourse. In this case, the computer mainly assists the analyst in ascribing a set of predefined categories to each relevant token expression in the text. This is what is done using content analysis programs such as NU*DIST or Atlas, for instance. But this procedure is time- and energy-consuming when the corpus is large. In CARAT, the procedure is reversed. The computer is used as a heuristic tool for discovering the possible categories themselves and the links between them.

In our research, this procedure was explored using an anthropological corpus composed of transcripts of interviews done in a Native-Canadian community: the Innus of Quebec.

Two types of analysis were developed. The first one was a classification analysis. It consisted in extracting the classes of segments specific to a particular chosen theme. This procedure is similar to the previous one and it allows one to discover unseen semantic relations between the expressions of the segments.
The second type is an analysis by “attractor” categories in which one discovers the expressions around which many others tend to polarize, that is, to form an “attractor theme.”

Let us now briefly illustrate this methodology. The first step involves choosing some particular classes of segments that seem to express some particular theme. This theme will be the attractor. All the terms contained in these classes are then extracted and organized in decreasing order of frequency for each class. Hence, some terms will be more important or relevant than others in the classes.

In a second step, all the “winning expressions” are then translated into a set of categories pertaining to a thesaurus. For instance, the expressions pot, marijuana, cannabis, etc., are transformed into the common thesaurus category [DRUG]. This transformation allows the analysis to work on the categorical labels instead of the words themselves.

In a third step, just as in the thematic analysis presented above, relations are then discovered with a neural net (ART) and identified though an analysis of the lexicon of each thematic class that relates to the chosen main theme or attractor. In the following example, we show the results in a graph (Figure 4) that reveals the semantic net for the attractor [DRUG]. And the arrows in the graph indicate a textual proximity relation that was discovered between the thematic concepts in the corpus.

The whole graph (Figure 4) represents how the main theme [DRUG] relates to other subthemes. For instance, starting from the top, the [DRUG] theme relates to young men (HJ), which in turn relates to it as a negative vital process (PV–) and a negative event (E–), etc.

![Figure 4. Exploration of subthemes for the attractor [DRUG]. In this graph, each circle represents a particular theme, such as the following: HJ, Young Male Human; or HF, Young Female Human; R, Religion; PV, Vital Processes (positive + or negative −); G, Geography; ES, Emotions and Sentiments (positive + or negative −); A, Actions (positive + or negative −); DC, Cognitive Domain; E, Event (positive + or negative −); RS, Social Relations (positive + or negative −); T, Transformation.](image-url)
Translated into more simple terms, an anthropologist sees in this network the following thematic structure, which is his own interpretation: in the Innu culture, relative to the question of drugs, the elder (the grandmother) (*Mamit Innuat*) uses Catholic religious rituals (prayers and pilgrimages) to influence the behavior of the young members of the community so that they will stop using drugs and so that they will not require the services, methods, or vocabulary of the social services implemented by the Quebec government in the aboriginal community.

As in thematic analysis, the computer does not produce this interpretation. It only produces some structuring of the textual data that makes important regularities in the corpus more obvious; the interpreter needs to pay attention to these features. The interpretation is based on these regularities.

6. The computer design: SATIM

As one can see, classification and categorization techniques can be successfully used in CARAT. By its nature, such a methodology requires a very flexible computer platform, because the various parameters of the methodology and the evaluation procedure cannot easily be defined and applied systematically. In order to execute this research program, all the preceding strategies must be performed on a very flexible computer platform. Only a few such platforms exist that are specially dedicated to the humanities and social sciences. We present here a brief review of our own in-house program called SATIM, which is a computer platform for high-level design of computer-assisted information processing. Its main function is to help design, produce, and experiment with complex processing chains on various types of information (text, multimedia, etc.). More specifically, it offers three levels of construction specialized for either a software developer, a researcher, or an end user. We can conceptualize the SATIM platform as presenting three levels of architecture called the *workshop*, the *laboratory*, and the *application*.

6.1. The workshop

This first level of the platform is an incubator and a repository for the various modules (functions) that are to be used in building analysis chains. This workshop deposits,
integrates, and manages various modules that a user may build or even acquire from various sources (if they are compatible). Hence, it serves as an integrated repository of various tools for computer text analysis (lemmatizers, parsers, classifiers, etc). These modules are autonomous and independent. They may even have been built in different programming languages. This is a condition for maintaining and updating the workshop.

This SATIM workshop is in fact a special type of toolbox. It is not so much a set of modules (functions) but a set of tools for working on the modules. These tools allow them to communicate their input and output in light of a specific research objective or kind of processing.

SATIM relies on three computer programming approaches: object-oriented design, multi-agent design, and, most of all, combinatorial functional design, as presented above. In its current form, SATIM is really a huge relational database managing a variety of existing modules (lemmatizers, matrix builders, classifiers, statistical packages, etc.). For the time being, this workshop is accessible only to expert programmers and its design is part of an ongoing software engineering project.

6.2. The laboratory

The second level of the SATIM platform offers CARAT researchers tools to explore, in a transparent manner, various analysis chains (built from the toolbox and repository). At this level, the user does not need to be a computer programmer, but he must be an expert in a specific CARAT application. Through an ergonomic interface, the user chooses the modules he wishes to see functioning in one or other of the analysis chains. He is assured that these chains are syntactically well formed, that is, they form a complex algorithm. Their semantics (i.e., their meaning) depends specifically on the task in question. That is, even though the modules may be combined in a well-formed manner, they are not necessarily semantically relevant for a particular task.

Two analysis chains have been built: Numexco and Grammexco. These two chains are classification chains for texts. One works with words as basic units of information, the other with \(n\)-grams. Other chains are being built as the research goes on (Indexico for indexing, Ontologico for ontology maintenance, Thematico for thematic analysis, and Categorico for automatic categorization of texts).

6.3. Applications

The third level of the SATIM architecture is for the end user. If, after many tests and experiments, a particular successful chain is finally accepted, it can be wrapped up as an autonomous application, which is transparent to the end user and where only certain parameters are modifiable. It may have its own interface. And if a particular chain does not fit a specific goal and modules have to be changed, one must go back in the laboratory and develop new chains.
7. Conclusion

In this Chapter, we have shown how computer techniques for classification and categorization can be used and applied in the field of CARATs. These techniques can diminish the burden for many researchers in the field of social sciences and humanities, and even in various textual information technologies, when they need to go deeper into the content of these texts in order to extract themes and organize possible navigation paths in them. In our opinion, it is only when more of these types of techniques have been transformed into ergonomic systems that the social sciences and humanities will really adopt them. For the moment, there is still much more research to be done in order to better understand and model the process of reading and analysis of text.

References

Eco, U. (1965), L’œuvre ouverte (Seuil, Paris).
Fish, S. (1980), Is There a Text in This Class? The Authority of Interpretative Communities (Harvard University Press, Cambridge, MA).


Schleiermacher, F. (1808–1809/1987), Hermeneutique (Cerf/PUL, Paris/Lille).

Chapter 44

GRAPH MATCHING, SYSTEM DESIGN AND KNOWLEDGE MODELING

GUY W. MINEAU

Université Laval

Contents

Abstract 979
1. Introduction 980
2. Knowledge represented as graph structures 982
3. Learning heuristic knowledge 984
4. Viability conditions 985
5. The complexity of learning 986
6. Categorization of knowledge in layers 988
7. Conclusion 989
References 989

Abstract

Modeling is a highly cognitive task. When it comes to the modeling of a software system, there exist guidelines regrouped in methodologies, but there is no precise algorithm to follow. Modeling is still an art, even for producing a computer model. This chapter shows that modeling complex systems has to take some practical issues into consideration, since the usefulness of the resulting model may be restricted by its subsequent use. In this chapter, I show that the complexity behind knowledge-based systems, particularly the graph-matching problem that limits the inferences that the software system can make, requires that the system designer questions the modeling activities themselves before engaging in the modeling phase of the system. Therefore, this chapter sets out to illustrate a not-so-obvious relationship between a technical aspect of knowledge-based technology – the graph-matching problem – and the modeling activities required to produce such a system.
1. Introduction

Software systems are becoming more and more sophisticated; they include tons of functionalities, some of which requiring concurrent processing, sensory data interpretation, detailed user modeling, interconnection to a network of computers that provide them with online services, etc. Furthermore, a quest toward autonomous systems is currently driving the software industry into what will be the next trend in software system development: multiagent systems [Wooldridge (2002)]. The economical advantages of having software systems that are more reliable, more autonomous, and offer complete and integrated functionalities are immediately relevant to their market value: they provide access to a broader market share and thus, to additional revenues.

In order for these systems to fulfill their purpose, they must model the world with which they interact with sufficient complexity. The design of complex models calls for knowledge-based technology. It is predictable that all complex software systems will eventually be designed as knowledge-based systems.

According to the CommonKADS methodology used for developing such systems [Schreiber et al. (2000)], many different types of knowledge are required in the making of a knowledge-based system: domain, task, and expert knowledge constitute just a few examples of these. Since knowledge is the key element of these systems, it will need to be acquired and be reasoned about, for the system to be able to make decisions and carry out the tasks for which it was designed.

In order to acquire knowledge and represent it within a computer, a knowledge-modeling language is required. This language is crucial to the knowledge-modeling process. It should have the following characteristics. First, it should have the right amount of expressiveness. The expressiveness of the language constrains what can and cannot be represented, and therefore, it impacts on the capability of the system to deliver the sought-for services to its users. Of course, its expressiveness should be such that in the end, the model of the reality that it allows to be represented is rich enough for the system to make the right inferences and carry out the right actions. On the other hand, if it is too expressive, it may result in a model whose complexity is such that system performance, and even decidability, is seriously jeopardized. Therefore, pertaining to the completeness of this representation language, a trade-off between overexpressiveness and underexpressiveness needs to be considered during the modeling phase.

Second, the knowledge-modeling language should be designer friendly, so that its use is not in itself a difficult hurdle to overcome, nor is it perceived to be one. A complex knowledge-modeling language certainly offers more expressiveness, but it comes with a higher level of difficulty. “Modeling,” as the term itself describes, addresses the problem of creating a model of some reality. The process of creating a model results in some encoding, in a particular modeling language, of some approximation of a reality. For example, if \( L \) is the modeling language and \( \lambda^* \) is the sets of all sentences in this language, then modeling a domain \( r \) can be seen as mapping bits and pieces of \( r \) to elements (sentences) of \( \lambda^* \), or in other words, the modeling activity will result in a set \( m \) such that \( m \subseteq \lambda^* \), where \( m \) hopefully describes \( r \) in sufficient detail for the system to make proper inferences and take
proper actions. Therefore, the modeling process requires an in-depth knowledge of $\lambda$ and of its modeling capabilities, and therefore, of $\lambda^*$. The mental model of $\lambda$ itself, required by the system designer in order to use $\lambda$, should be fairly easy to build and use.

Third, the language should be unambiguous. Knowing that the elements of $\lambda^*$ will be used in deterministic processing (by a computer program), no ambiguity should be associated with the use of $\lambda$. In other words, if $l_i \in \lambda^*$ and $l_j \in \lambda^*$ such that $l_i \neq l_j$ (syntactically), then $l_i$ and $l_j$ should carry different semantics, i.e., they should describe different phenomena of $r$. If $l$ is a formal knowledge-representation language, $l_i$ and $l_j$ certainly carry different semantics as most knowledge-representation languages provide canonicity of representation. The system designer, however, must interpret $l_i$ and $l_j$ correctly, not confusing them for one another, which, otherwise, would result in using one or the other interchangeably, producing erroneous inferences. Easy mapping from each (or more realistically, most) $l_i$ in $\lambda^*$ to the reality that it models is required. The designer will perceive that property as ease of interpretation, which would result in categorizing $\lambda$ as being designer friendly even if $\lambda$ needs to be a formal representation language.

So, in summary, the designer must precisely assess the appropriate level of knowledge representation that the application will need to be able to achieve its assigned tasks. Within the boundaries set by that important constraint, the choice of an appropriate $\lambda$ is made, so that the one chosen is as designer friendly and unambiguous as possible, especially in light of the complex development cycle of a software system, where many individuals will be involved in different steps of system development that range from system analysis and design to testing and reiterating.

This is why graphs are so popular in modeling tasks.

(a) They are pictures and a picture is worth a thousand words, provided that its complexity is somewhat restricted, so that it remains simple to interpret.

(b) Their nature provides similarity with other familiar learning tools that were and might still be used in an everyday basis: reading a map, playing hopscotch, going through a labyrinth, driving through a network of streets and highways, etc., and for system designers: entity relationship (ER) diagrams, Unified Modeling Language (UML) diagrams, state charts, Petri nets, flowcharts, Resource Description Framework (RDF), etc.

(c) They offer simple modeling tools: vertices, arcs, and labels.

(d) They offer natural focal points in some basic elements of the model (vertices) upon which complex modeling is carried out (using arcs and eventually, layers). Focusing permits the zooming-in on parts of the model under development, and, therefore, the isolation of more complex parts of the model to be dealt with separately or at a later stage, and, consequently, the implementation of an efficient iterative approach to model design – as proven by dozens of graphical modeling formalisms that exist not only in the literature but also in the market place. Handling complexity progressively is definitely one way to simplify system design.

(e) By projecting graph structures onto one another, the reasoning path (the inferences) can be visualized, which would help to produce the associated explanations, whenever needed.
So, for the sake of argument, in the following text, I chose $\lambda$ to be the conceptual graph formalism [Sowa (1984)]. My rationale is that it was demonstrated in the past that the conceptual graph (CG) notation carries the expressiveness of first-order logic, while supersets of the notation allow the modeling of more complex realities: behavior, implication rules, constraints, modalities, fuzzy statements, etc. Therefore, a bidirectional mapping between the CG notation and many other modeling languages is possible, making the conclusions that will be drawn below from the use of the CG formalism also true of other notations.

2. Knowledge represented as graph structures

Inferring new knowledge from two graph structures, $g_1$ and $g_2$, implies a comparison of these graphs in order to find whether they carry complementary knowledge. If so, new knowledge can be inferred through a chain of inference. Figures 1 and 2 give an example of (conceptual) graphs whose comparison results in an inference chain that produces other graph (knowledge) structures. This comparison of graphs is called the projection of a graph onto another [Mugnier (1995)], and its complexity is bound to that of the really difficult, isomorphism problem in graph theory [Garey and Johnson (1990)].

Comparing graphs is a complex process. The human mind seems to find quite easily the obviously relevant parts of two graphs to be matched – when the graphs are simple – but a computer program has to try matching the graphs exhaustively, on every possibility. Notwithstanding the decidability issue associated with knowledge-based systems, this exhaustive search may result in a system that does not find an answer within the allowed (or targeted) response time. Consequently, much research in graph (and pattern) matching explored at length heuristic (trial and error) approaches to the graph-matching problem. In many cases, simplifying assumptions must be made in order to constrain the complexity of the task.

Fig. 1(a). Graph $g_1$: All students are enrolled in some program. (b). Graph $g_2$: A student, Sophie, likes biology. (c). Graph $g_3$: A student, Sophie, is enrolled in some program.
Some applications favor the matching of graphs from pre-identified vertices ("people" with "people" for example, favoring a semantic proximity between vertices). Other applications restrict the shape of the graphs that they handle. For instance, many restrict the number of circuits (or loops) in a graph, and, as a result, have to deal mostly with tree-shaped graphs; others restrict the number of variables that can appear in a graph, limiting the number of potential matches between vertices. In any case, all simplifying assumptions either (a) restrict the representation space so that the complexity inherent to the representation structure is limited (as with tree-shaped graphs), or (b) restrict the search space based on some a priori knowledge of the needs and particularities of the application domain (using a heuristic-based search).

But generally speaking, the expressivity of the representation language is dictated by the needs of the application domain, as mentioned earlier. Therefore, restricting the representation space may not be an option. In addition, the availability, soundness, and completeness of a set of heuristics with regard to the implementation of a heuristic-based projection algorithm for a particular application domain is mostly based on extensive experience and knowledge of that domain. That leaves system developers with fundamental questions to answer; the replies have a bearing on the usefulness of the system under development. Since the usefulness of a software system in the marketplace is measured in terms of the revenues that are generated directly or indirectly (through reduced expenses), key elements in the decision-making process pertaining to the development of such a system, are the answers to these questions:

1. Are such heuristic knowledge sources available?
2. If not, could one generate heuristic knowledge sources?
3. If so, how can one rely on heuristic knowledge sources?

---

Fig. 2(a). Graph $g_4$: If some student likes a particular subject, then s/he is enrolled in a program related to that subject. (b). Graph $g_5$: A student, Sophie, is enrolled in a program related to biology.
It is not easy to answer the first question: it all depends on the availability of relevant knowledge sources. In some instances, reliable knowledge sources about the application domain may be available from experts or documented cases. In these situations, the next issue would be to determine how to acquire and transfer available knowledge into a computer program. There are methodologies that address these issues, like CommonKADS [Schreiber et al. (2000)].

But what if knowledge sources are not available? This question will be addressed in the next section; it is basically the whole point that this chapter is trying to make. I believe that under some circumstances, heuristic knowledge can be learned (inferred) from example cases. In effect, I believe that machine learning theory [Mitchell (1997)] could be applied to generate that knowledge and Section 3, which follows, gives an outline of the learning model that would be required to do so. Provided that the conditions needed to generate that heuristic knowledge are not complicated to attain (see Section 4), I suggest that the answer to the first question could be: “we can generate this knowledge based on past experience.” That way, we could produce our own knowledge sources in situations where such sources are not directly available.

The third question is a fundamental one, since the soundness (and completeness) of the system may be at stake. One must have the assurance that the heuristics that will guide the projection (matching) algorithm are in line with the needs and particularities of the application domain. If not, the consequences of having a system that would not behave soundly (i.e., would have sporadic faulty behaviour), or that would not take actions in all normally required situations, should be carefully assessed. The soundness and completeness of a set of heuristic knowledge is quite dependent on the reliability of the sources where these heuristics come from. Under my proposed model that follows, the confidence that we have in the set of heuristics that would be generated may be statistically assessed, as machine-learning theory deals with these aspects.

3. Learning heuristic knowledge

Since graph matching is at the heart of the inference processes that support the behavior of knowledge-based systems, and since it is rarely the case that such a system benefits from expert knowledge that can dictate how graph matching should be handled, I propose, in this section, a learning model that would infer these heuristics.

Learning is based on past experience – some good, some bad. Behavior is reinforced accordingly: one tries to relive good experiences while avoiding bad experiences. Experiences can therefore be characterized as either positive or negative situations. So one has to understand why a particular situation is considered positive while some other may be viewed as negative. Learning is about exploring both types of situations, comparing them, and finding the common characteristics of each type, so that one could predict the type of a newly encountered situation and thus adapt its behavior accordingly.

With our aim of learning heuristic knowledge about how to match graphs, the situations composing our set of experiences $E$ are possible matches between pairs of graphs.
So each situation $s_i \in E$ will be described as: $<g_1, g_2, g_3>$, where $g_1$ and $g_2$ are two graphs, and $g_3$, one of their possible matches. ( Needless to say, $g_1$ and $g_2$ are associated with many situations.) Furthermore, feedback being the key factor in reinforcement mechanisms, let us define the feedback function $f: E \rightarrow [-1,1]$ for all $s \in E$, with $f(s) = 1$ as being the most positive rating. Since the image of $f$ is a linearly ordered domain, for $E' \subseteq E$, a ranking of all situations in $E'$ is possible through the use of $f$, so we know which situations should be favored. Finally, since $g_3$ is a match between $g_1$ and $g_2$, it can be demonstrated that $g_3$ is a common generalization of both $g_1$ and $g_2$ [Sowa (1984)]. Generalizations are generated from the application of generalization operators, which are of three types [Sowa (1984)]: (1) turning a constant (an individual) into a variable (that represents a set of possible individuals), (2) generalizing the type of some concept (represented by a vertex) (e.g., a student being a person, one could use the latter more inclusive, term rather than the former term), (3) dropping part of the graph structure, which is equivalent to forgetting some information (which does not invalidate the remaining information). In order to produce $g_3$, a sequence of these three operators was applied on both $g_1$ and $g_2$; it can be directly associated with $f(s)$. So, we now have all the information we need in $E$ to compare the various situations in order to find out which operators applied to which types of graphs resulted in good matches.

However, the applicability of our learning model depends on the existence of a set of situations $E'$ (that does not need to be exhaustive, so $E' \subseteq E$), and on a feedback function $f$ (that assesses the usefulness of the resulting match), as will be discussed in Section 4. Also, complexity issues pertaining to the learning algorithm itself may jeopardize the feasibility of the proposed approach (as explained in Section 5). The rest of this chapter discusses these two issues which greatly affect the feasibility of inferring heuristic knowledge pertaining to the matching of knowledge structures.

4. Viability conditions

As mentioned earlier, the existence of rated examples is a sine qua non condition for any learning system to operate. Either some experts have provided some examples and have manually rated them, or there exists a rating (feedback) function that can be applied on newly (manually or automatically) created examples. In the first case, experts have to be available. For our purpose, we need experts on graph (knowledge structure) matching in the target application domain. Such resources are rare. Experts do not reason about their own cognitive processes. Domain experts may not know how they compare and match different knowledge structures. Furthermore, the knowledge representation language in which matching needs to be carried out, $\lambda$, will most probably be totally alien to the domain experts. Therefore, some automatic way of producing and testing examples would be best.

In some cases, when the application domain may be described by $\lambda$ in such a way that syntactically correct sentences are also semantically correct, then by handling the construct operators of $\lambda$, as in any formal language that has a well-defined grammar, a computer program can generate sentences in $\lambda^*$. So there exist cases where $E'$ can be
generated. For instance, let us take a database application. The application domain is normally described (modeled) using some graphical formalism (like ER or UML diagrams). There exists a domain ontology. The data stored in the database describe the actual state of the domain so modeled. And Structured Query Language (SQL) queries that are sent to the database respect a strict grammar. Based on the database model (a graphical model), some computer program could generate SQL queries (which can also be represented as graph structures) and evaluate them against the content of the database. So there are applications for which \( E' \) could be generated automatically.

But then comes the problem of rating the automatically produced examples. Experts are good at evaluating cases. They know whether and how much these cases are relevant to the query. The problem with involving experts is the following. First, they must agree on the usefulness of the cases (results of the system). In order to statistically validate their feedback, one usually needs to involve an odd number of experts – the larger the number, the better. This comes with a high price tag, since experts are difficult people to reach, and are expensive people to hire. Furthermore, their time is limited, which limits the size of \( E' \) (since every example in \( E' \) needs to be rated), which therefore limits the statistical validity of the inferred heuristics.

In some cases, when the rating function is based on some automatic assessment of the results, feedback may be automated. In conjunction with the automatic generation of examples, that allows the production of a much more sizeable \( E' \), which would entail a much more statistically significant set of inferred heuristics. In our example, one could assess the usefulness of the result of an SQL query by some numerical measurement like the number of tuples in the answer. For many information-retrieval purposes where maximizing that number is sought (i.e., maximizing recall), this assessment can be achieved automatically. In brief, for all applications where the assessment of the results can be automated, a large \( E' \) can be produced and provided to a learning algorithm that would infer statistically significant heuristics.

In summary, the applicability of the learning model proposed in Section 3 depends on the availability of \( E' \). Either the human resources are available to produce a \( E' \) of sufficient size, or the nature of the examples and of their rating function allows for some automation of this process.

5. The complexity of learning

The set of experiences \( E \) is defined over all possible situations \( <g_1, g_2, g_3> \), where each \( g_i \in \lambda^* \). Therefore, the learning algorithm can be defined as a function \( l: \lambda^* \times \lambda^* \times \lambda^* \rightarrow \lambda^* \), so that its output can classify any new situation according to its perceived value (according to the rating function \( f \)). That is, for each newly encountered situation, a preferred match \( g_j \), a generalization of \( g_1 \) and \( g_2 \), should be identified by the learning module based on the analysis of the situations in \( E \) according to the rating function \( f \). So the search space of \( l \) is large; it contains \( |\lambda^*|^3 \) elements (at most), where \( |\lambda^*| \) is the size (cardinality) of \( \lambda^* \), the number of sentences that can be generated using \( \lambda \).
It is obvious to see that the complexity of $I$ depends on two conditions: (1) the size of $\lambda^*$ and (2) the search algorithm that explores $E$. For the first condition, it is the representation space that needs to be kept as small as possible. As mentioned before, the system designer needs to choose a $\lambda$ that brings just enough expressivity in order for the important aspects of the reality $r$ to be modeled, to be captured in a model $m$, so that the system can achieve the tasks for which it is designed. We now see why unnecessary complexity of $\lambda$ may bring additional complexity to $m$, and may be fatal to the system if the associated complexity refrains the learning model from efficiently learning how to match the knowledge structures encoded using $\lambda$.

For the second condition, it is the size of the search space that is actually searched that one tends to minimize. In effect, one can imagine very naive search algorithms that would sweep $E$ sequentially (in cases where $\lambda^*$ is finite), trying to match every possible pair of graphs in $\lambda^* \times \lambda^*$. Some other algorithms may be more sophisticated and use various relationships between the elements of $\lambda^*$ to guide their search through $\lambda^* \times \lambda^*$. In effect, for all pairs of elements $g_i$ and $g_j$ in $\lambda^*$, one of the following (comparative generalization) relations holds. We either have: (1) $g_i \prec g_j$ (which means that $g_i$ is less general, or more specific, than $g_j$), (2) $g_i \succ g_j$, (3) $g_i = g_j$ (g_i is as general as g_j), or (4) $g_i \neq g_j$ (g_i is not comparable to g_j). That is, there is a partial order of generality defined over the elements of $\lambda^*$; and the search algorithm may use this partial order to avoid losing time in subspaces of $\lambda^* \times \lambda^*$ where, based on previously obtained results, there is no point in exploring some more. For instance, one could imagine that if $g_i \neq g_j$, then any of their specializations will not be comparable as well; so there is no need to consider situations starting with $<g_k, g_m, \ldots>$ with $g_k < g_i$ and $g_m < g_j$. Of course, learning algorithms use such knowledge and many other techniques to optimize their search through the learning space, even using probabilities, random exploration, etc. Machine learning theory, and more specifically, computational learning theory [Kearns and Vazirani (1994)] deal with these aspects in length.

As the choice of a learning algorithm should be left with an expert in machine-learning theory, the system designer still has the responsibility to choose a particular $\lambda$, appropriate to build the required model $m$. It may be the case that the expressivity needed of $\lambda$ produces a $\lambda^*$ that is very large – too large to efficiently have a learning model infer the appropriate heuristics needed to eventually match the knowledge structures encoded using $\lambda$, and whose matching is important in order for the system to achieve its tasks. So the question that remains is: what can one do when the search space, though as minimal as can be, is still too large for efficient learning.

Of course, there are learning techniques that deal with sparse search spaces that could be applied here. But more closely related to the task of system design, as is the point of this chapter, the system designer has to see whether the representation space can be modeled in layers, each layer adding additional features to $\lambda$, thus providing the necessary elements to model more complex aspects of $r$. So we would end up having a series of modeling languages $\lambda_i$ (for $i \geq 0$) where $\lambda_i \subset \lambda_{i+1}$, such that $\lambda_{i+1}$ brings at least one additional feature not appearing in $\lambda_i$, making it more complex than $\lambda_i$. The idea behind that principle is to segment the representation space in different layers of
complexity. The learning algorithm could infer heuristics on how to match the knowledge structures of layer \( i \) using \( \lambda_i^* \) as its experience set \( E \). So learning could occur layer by layer, until the complexity of the task becomes such that learning is no longer feasible. By dividing the complexity of the learning task in layers, one could hope to produce reliable heuristics for the lower layers, providing reliable heuristics to the matching algorithm of the knowledge structures described in these layers. That does not, however, solve the problem of reliably matching the knowledge structures of the higher layers.

6. Categorization of knowledge in layers

As knowledge modeling usually starts with the description of generic frameworks, called “templates,” from which more precise knowledge elicitation then proceeds, this idea of building knowledge structures in layers of complexity must be taken into account in this categorization process. In effect, not only should the templates be related to different subdomains of the domain ontology – providing for a broad classification scheme of the knowledge structures acquired later on – but they should also be related to the various features of \( l \) that they need to use in order to encode the appropriate knowledge.

For instance, let us consider the graphs of Figures 1(b) and 2(a). One can see that they require different features of the conceptual graph notation, our representation language. Figure 1(b) displays a graph that can be associated to proposition logic, as no variable is used in the graph: all objects that it contains are identified by constants. On the other hand, the graph of Figure 2(a) is more complex: it uses variables (bound by their indices), and is in fact a second-order graph where the if-then structure can be seen. Figure 2(a) represents a meta-level graph, a graph whose purpose is to define causal relations between first-order graphs.

Consequently, in the development of a knowledge-based system in layers, the categorization process must have at least two dimensions: the overall topic covered by the graph (both graphs of Figures 1(b) and 2(a) provide information about students), and their degree of complexity (the graphs of Figures 1(b) and 2(a) belong to two different categories: they are a prepositional first-order graph, and a meta-level second-order graph, respectively).

These distinctions must be made during the early stages of the modeling of the system. The results obtained on the complexity dimension will guide the system engineer in choosing, developing, and fine-tuning the appropriate operators that will be devised to specifically handle the knowledge structures defined at the same level of complexity, as explained earlier. Therefore, this bidimensional categorization is essential to the development of the system, and should be integrated to the various modeling processes that the system must undergo. This is, of course, another argument that states that system modeling cannot occur without a good understanding of \( \lambda \), the representation language that will eventually be used to encode the various knowledge structures identified as important to the system, tying system modeling and system development together, as is the main point of this article.
7. Conclusion

In summary, choosing a modeling language $\lambda$ is a crucial task. The subsequent modeling activities will depend on it: $\lambda$ should be designer friendly, as explained earlier. For the design of complex systems, especially when knowledge-based technology is required, it is foreseen that the complexity of the modeling task will require a complex modeling language. However, as was shown earlier, the complexity of the modeling language should be kept to a minimum as it impacts on the subsequent ability of the system to learn how to match the knowledge structures that it describes, since it is generally the case that the knowledge required to write relevant and efficient graph-matching algorithms is usually not available, though vital, to the functioning of such a system.

Since the complexity of $\lambda$ is dictated by the needs of the application domain, the system designer may need to structure $\lambda$ in different layers of complexity in order to allow a learning algorithm to learn heuristics that address the matching of knowledge structures of different levels of complexity. One could hope that for simple cases, the inferred heuristics would be appropriate, covering altogether a broad range of the cases that the system would normally have to handle, though being conscious that there will most probably remain cases that will be left without such heuristic knowledge to help the matching (projection) algorithm.

The point of this chapter was to show that the modeling activities required in the development of complex systems have to take into account the complexity issues pertaining to the functioning of that system. I aimed to show that modeling a software system – though a highly cognitive task – is a task that, though at the front end of the development cycle of a system, cannot be isolated from the other phases of system development. In addition, I aimed to show the associated complexity of the running system, particularly in the case of knowledge-based systems. I hope that we have stressed the importance for any system designer to foresee the eventual shortcomings of the system under development, and how the modeling activities should take into consideration part of these shortcomings.

References

This page intentionally left blank
PART 10

THE NATURALIZATION OF CATEGORIES
Abstract

Nominalism is the doctrine that only singular things really exist. Species, sets, common natures, general properties, shared attributes, and so on, are thus taken to have no mind-independent reality of their own. The aim of this chapter is to explain how this philosophical notion bears upon the theory of concepts as mental representations. The basic principle is that generality is not a feature of the things in the world, but of our representations of them. Examples are taken from the recent literature on concept theory, where confusion reign. The main motivations for nominalism are then briefly recalled. And, most importantly, it is shown how the nominalistic approach can be cashed out, in the field of concept theory, in terms of two main constraints: (i) anything that is supposed to be represented by a concept should be a singular thing; (ii) anything that is supposed to occur as a mental representation should also be a singular thing itself.
1. Nominalism

Nominalism, as I use the label, is the doctrine that only singular things exist in the world. Species and genera, according to this approach, do not enjoy any existence of their own: there is no “horse in general” or “animal in general”, no common properties such as redness or triangularity, no sets as distinct from their members, and, in radical versions of nominalism, no abstract entities either, such as numbers or Fregean senses that are simultaneously graspable by different minds. Only horses (in the plural), particular animals, trees, tables and chairs, human beings, the Atlantic Ocean, the Moon, and the like, populate the Universe. As the American philosopher Nelson Goodman once put it: “For me, as a nominalist, the world is a world of individuals” [Goodman (1956, p. 13) Goodman (1972, p. 155)]. What I want to explain here is how this philosophical outlook, if we are to accept it, bears upon the theory of concepts.

The term “concept”, of course, has a wide variety of meanings. In philosophy, it is often used, on the heels of Gottlob Frege, to stand for some purported abstract entities which are taken to be the meanings of words and the components of propositions, understood as the objects of thought. Concepts, in this view, are supposed to be mind-independent: although they do not exist in space-time, they are seen as full-fledged elements of objective reality, of which they are said to constitute the “Third Realm” (in addition to the material and the psychological domains). This is not the sense of the term “concept” I will be using here. For one thing, there are reasons to be skeptical of the usefulness of postulating such special entities, even as theoretical posits. But the main point, at this stage, is simply that what I will be interested in is the theory of mental categorization. And there exists, in this context, a different but equally widespread acceptation for “concept”, according to which it stands for psychologically realized mental representations that are both general and combinable with each other into sentence-like units. Stephen Laurence and Eric Margolis, for instance, straightforwardly announce in the introduction to their important collection of readings on concepts that they “will take concepts to be subpropositional mental representations” [Laurence and Margolis (1999)p. 4]. This is the sense I will be using, and it is also the one that is favored in recent discussions within cognitive sciences, quite independently of whether they are nominalistically oriented or not [see, for example, Keil (1989), Medin (1989), Komatsu (1992), Fodor (1998), Murphy (2002), and a host of others].

Let me insist that what I call nominalism is not the thesis that all concepts thus understood are language-dependent. Admittedly there is such a use of the label “nominalism” in the context of concept theory. In a recent book, Wayne Davis, for example, severely criticizes what he calls the “nominalistic theory of concepts” as being “wide off the mark” [Davis (2003), p. 428]. But what he understands by “the nominalistic theory of concepts” turns out to be the thesis that all mental concepts reduce to certain dispositions to use words. This is not what I will be talking about. As a nominalist in my sense, I have no quarrel with the suggestion that dogs or gorillas can be said to have concepts even though they have no language in any strong sense1. Even in human

---

1 Such a suggestion is developed, for example, in the works of the French philosopher Joëlle Proust on the animal mind [e.g., Proust (1997, 2003)].
beings, as Jerry Fodor has frequently argued, the possession of concepts seems to be a prerequisite for learning a language in the first place [Fodor (1975)]. Our own concepts, then, cannot all be language-dependent.

Nominalism, in the sense I favor, is an ontological doctrine. It has to do with what the basic furniture of the world is. It has been traditionally labeled as “nominalism” – especially from the fifteenth century on – because it was taken to be one of the possible answers to the so-called problem of universals: what sort of things are species and genera? This answer, as put forward by Peter Abelard in the twelfth century and William of Ockham and his followers in the fourteenth century, was that universals are names (nomina) rather than external things. But “names”, in the fourteenth century at least, were not exclusively supposed to be conventional units belonging to such external languages as English, French, or Latin. Names, in this vocabulary, were primarily thought of as the basic representational units of the language of thought – the oratio mentalis – units which these authors also called conceptus. The late medieval nominalistic thesis, then, was that generality is not a special mode of being, but a semantic or representational feature of certain mental units. This is still pretty much the core of today’s nominalism as I understand it.

This doctrine, of course, is not necessarily naturalistic. The medieval philosophers after all, whether they were realists or nominalists, all firmly believed in the existence of a supernatural realm populated by purely spiritual beings such as God and the angels. Yet it easily lends itself to being harmonized with a generally naturalistic outlook insofar as it refuses to countenance non-spatiotemporal abstract entities.

2. Ockham’s cleaver

William of Ockham’s name is generally linked in recent literature with the principle of parsimony known as “Ockham’s Razor”. This is the principle that entities should not be multiplied without necessity in explanations. It is a healthy principle and Ockham did indeed make use of it in his writings – but not as centrally as is often thought. The so-called Razor principle did not originate with him; it was commonly accepted in the decades before he began to write. And it is not, I must insist, one of his main arguments in favor of nominalism. On the other hand, there is another principle which, although he did not explicitly formulate it, is more central to Ockham’s nominalism, and much more important, I believe, even for us: it is the principle that the American medievalist John Bolker has aptly labeled “Ockham’s Cleaver”. In its mildest form, what this principle says is that we should be cautious not to take as features of the things signified the
features of the signs that signify them. In other words, we should not conflate representational features with ontological ones [Boler (1985)].

It is easy to see how the nominalistic doctrine is supposed to be – for those who accept it – an application of this principle to generality. Generality, the doctrine says, is a feature of representations qua representations and it should not be mistaken for a feature of the things represented. But, whether or not they accept nominalism as a plausible ontological theory, the Cleaver principle should never be lost sight of by those who are professionally and theoretically interested in representations and signs, such as philosophers of language and mind, linguists, and psychologists. Neglecting the Cleaver principle is the direct source of a surprising amount of confusion, and I would be quite happy if my presentation of some nominalistic ideas in the present context helps to develop awareness of this crucial distinction between talking about representations (or signs) and talking about what is represented by these representations. Let me give a few examples, drawn from the recent literature, of how widespread such confusion is in the field.

My first example has to do with the distinction between concepts and categories in the relevant literature. Medin and Coley (1998,p. 404), for instance, provide the following useful characterization:

By concepts we mean a mental representation of a category serving multiple functions, one of which is to allow for the determination of whether or not something belongs to the class. A category refers to the set of entities picked out by the concept.

Prima facie, that is clear enough: a category is a set of entities, and a concept is a mental representation of such a set. A nominalist, of course, will spontaneously worry about whether or not such a characterization is supposed to convey a commitment to the real extramental existence of sets as distinct from their members, but let us leave this question aside for the time being, and simply admit that what Medin and Coley mean is that a concept jointly represents all the singular entities which belong to a certain category, rather than the set itself understood as a superentity. The generality of a concept in this view consists in its simultaneously representing several singular entities, rather than one single thing. Thus taken, Medin and Coley’s formulation nicely fulfills the requirement of Ockham’s Cleaver: a concept is a mental representation, and the corresponding category is the set of things that are represented by the concept. The vocabulary, here, traces a clear demarcation between the representation and the things represented.

This distinction, however, unfortunately tends to be blurred in much of the literature I am familiar with. Take the discussions that have been going on since the 1970s concerning the Prototype theory of concepts put forward by Eleanor Rosch. This theory holds that the most central component of a concept is a synthetic representation of what typical members of the corresponding category look like, and this prototype representation is said to be what subjects turn to in deciding whether a given instance does or does not belong to the corresponding category [Rosch (1975, 1978)]. So far, so good. The discussions surrounding this theory, however, have regularly conflated what is supposed to
be the characteristics of the relevant representations with those of the corresponding category. A central – and well-established – tenet of Prototype theory is that our concepts for concrete objects such as birds, fruits, weapons, or furniture routinely pick up some instances as better examples of the category than others: a robin is standardly taken (in North American culture, at least) to be a better example of a bird than a penguin is. And this accounts for well-documented typicality effects. The problem is that, because of this, the theory has often been taken to entail that the corresponding categories should have fuzzy boundaries, such that belonging or not belonging to such a category would be a matter of degree: this is what is called graded membership. Reisberg, for example, in his well-known textbook on the psychology of cognition, writes the following: “The graded membership notion, therefore, is a striking and important implication of the prototype view” [Reisberg (1997)p. 311]. This means, apparently, that according to the Prototype view of concepts, penguins are birds to a lesser degree than robins are! Or at least that most North Americans believe they are. But quite obviously, this is not so!

Three things need to be sharply distinguished here:

(i) which individuals (or groups of individuals) best exemplify a concept;
(ii) which individuals actually belong to the corresponding category;
(iii) which individuals are believed to belong to the category.

The distinction between (i) and (ii) is what the principle of Ockham’s Cleaver draws attention to in this case. The distinction between (i) and (iii), on the other hand, follows from the fact that people are normally able to distinguish between (i) and (ii) and do not generally believe that better examples of a given concept always have a higher degree of membership in the corresponding category. Some theoreticians, including Rosch herself at one point, and – by his own admission – such a respected scholar as George Lakoff [(Lakoff (1987)pp. 44–45), plus numerous others of course, have fallen headlong into this trap. And it has certainly created unnecessary difficulties in the literature.

The second example I will take of the dangers of neglecting the Cleaver principle has to do with the idea of an attribute or feature, which looms large in the debates of the last two decades in the field of categorization theory between the Prototype approach to concepts and the Exemplar theory. The latter holds that our mental representation of a certain category basically consists in the stocking of singular representations of salient instances we have experienced in the past. Our decision as to whether a certain newly encountered individual does or does not belong to the category is supposed to be a function of the similarities we notice between the new case and the previously stocked ones. Both Prototype theory and Exemplar theory are similarity-based: what we look for are similarities with the mentally constructed prototype according to the former, and with previously

4 For good summary reviews of typicality effects, see, among a host of others, Medin and Smith (1984) or Murphy (2002, p. 11–40).

5 May I insist that I am not claiming that there are no fuzzy categories at all. Obviously, there are (e.g., the category of small animals). The point is that graded membership in a category does not simply follow from the graded structure of the corresponding concept with respect to typicality.
encountered exemplars according to the latter. But how is similarity measured? Researchers differ as to the exact mathematical model to be used, but many agree that “[…] similarity is a function of the number of attributes (weighted for relevance or salience) shared and not shared between objects being compared” [Komatsu (1992), p. 514].

Even if one is not a nominalist, there is a problem here that we should be aware of: how are those “shared attributes” supposed to be individualized? This is often simply taken for granted: there is the color, the shape, the way the thing moves, and so on… Isn’t that clear enough? Well no, it isn’t clear enough! Komatsu, for example, in discussing these similarity-based theories of concepts, is led, like several other authors, to distinguish between “the accessible, easily perceptible attributes” – or “surface attributes” – such as color and shape, and the “less accessible, deeper attributes”, such as hidden causal powers in particular [Komatsu (1992), p. 514]. This strongly suggests that the relevant notion of a shared attribute is a little mysterious. And one might wonder: what is it that gives its unity to a shared attribute? And what does its “commonality” – the alleged fact that it is common to several things – consist of? A suspicion that comes to mind is that what is common in such situations is not a special sort of intrinsically general entity with a peculiar mode of existence which would allow it to be simultaneously present somehow within several singular instances, but a representation or a label; and that such a representation is common to several things only in the straightforward sense that it simultaneously represents all these things. If this is correct, counting the so-called shared attributes of singular things amounts to nothing but counting certain labels that apply to these different things. Commonality, in this view, turns out to be a feature of representations, not of the things represented. Once we become aware of this possibility, we begin to see the problems a little differently. We begin to suspect that similarity-based theories of concepts, as they are discussed today, presuppose the previous availability of certain general representations – let us call them “attribute representations”–which are simply left unexplained.

In a typical experiment on the concepts of [FRUIT] and [VEGETABLE], for instance, Edward and his collaborators developed a list of 26 attributes that could be considered as relevant among the subjects they worked with for categorizing a certain object as a fruit or a vegetable [Smith et al. (1988)]. These included color, texture, size, shape, smell, and taste, but also more abstract and relational properties such as: how it is eaten, how it is grown, nutritional value, typical consumer, cost, and so on. It is quite obvious that the latter, at any rate, are not immediately observable and are not individuated as such in direct perceptual experience. In what sense, then, do the 26 items belong to a single list? What is this list a list of? The most natural answer would seem to be that it is a list of predicates (or groups of predicates) with respect to which various singular instances can be compared in the categorization process. If so, the whole process depends on the previous availability of these predicates, which are themselves individualized as conceptual – or even linguistic – units. And what the researchers are actually counting when they calculate similarities are not special “shared” things out there in the world, but various representations or labels that subjects tend to apply to singular instances. In such a case, the use of Ockham’s Cleaver reveals
that similarity-based approaches to such “concrete” concepts as [FRUIT] or [VEGETABLE] crucially involve unexamined assumptions about another group of – more abstract – concepts, namely attribute representations. Whether we end up as nominalists or not, this is a kind of problem that we should not ignore.

My third example will be more specific. It has to do with a suggestion by the French scholar Jean-Pierre Desclés that a special theoretical idea should be introduced in the field of concept theory, that of the “typical object” [Desclés (1993)], which, Desclés contends, is regrettably absent from the recent psychological literature on concepts. Yet the way Desclés explains this idea in his 1993 paper on the subject nicely illustrates the sort of confusion I want to draw attention to. He defines the typical object for a given category as an object which is such that if anything belongs to the category, then it does (formally: the typical object for a category “C” is an object x such that if (for some y) Cy, then Cx). On the other hand, however, Desclés insists that the typical object for any given category is an “ideal entity”, “without any direct empirical correlate” [Desclés (1993), p. 151 (my translation)], and that being so typical, it lacks any precise determination:

The typical object is not a determinate instance of the category. In fact, the typical object for a property is the object which, as an objectual entity, “best” represents the property in all its generality and its determination would turn it into a “less pure” representative, with a diminished potentiality. [Desclés (1993), p. 152 (my translation)]

The typical object for the category [HUMAN BEING], for example, would have two eyes presumably, but neither blue eyes, nor brown eyes, nor black eyes, since any such determination would diminish its representativeness.

The problem, of course, is that nothing can belong to the category of human beings, let alone be a typical representative, and yet be so indeterminate. If x is a human being, as the typical human being should be, then x has eyes of a determinate color! Indeterminacy in such cases can only be a feature of representations, not of anything represented. Taken at their face value, Desclés’s explanations about the typical object bluntly conflate representations with represented things.

3. Motivations

I am not saying that nominalism is the only available way of avoiding the kinds of confusions illustrated in the previous section. My point so far has been that typical nominalistic preoccupations, even if they are not thoroughly endorsed, are especially helpful for the identification and clarification of the kind of issues I have raised, and for sharpening our awareness of the problems, in particular, those that have to do with the crucial distinction between talking about representations and talking about represented things. Yet nominalism has more to it than a mere consciousness-raising function. This is not the place to argue for it at length, but let me briefly recall, before returning to concept theory, some of the main considerations that prima facie speak in favor of a generally nominalistic outlook.
First and foremost, there are epistemological reasons. What we are aware of in our experience are singular entities. I meet with human beings, dogs, and trees, rather than with manhood, dogness, or treeness. I see red objects and round things rather than redness or roundness. Manhood as a nature or redness as a general property are not perceivable in themselves, it would seem. Any naturalistically minded approach to cognition, then, should, if it can, dispense with such hypothetical, unperceivable entities. This is especially true for those of us who are interested in a causal account of knowledge, since causality is – at least prima facie – a connection between singular things or events.

Second, the recourse to general and abstract entities does not usually have much explicative import. Participation in the universal redness, for instance, is supposed to be the metaphysical foundation for anything to be red. But saying that this shirt or that flower participates in – or exemplifies – redness seems to be nothing more than a contrived way of saying that it is red! With respect to the explicative value of such sentences, it is difficult to see how these realist formulations shed any light on what is going on, and the net result of introducing the supposedly foundational connections of participation or exemplification seems to be merely a loss in intelligibility. With reference to the two main varieties of realism that have historically been proposed (universals as separate entities, and universals as internal components of singular things), John Heil succinctly but sharply summarizes the point in his recent book:

If universals are Platonic entities instantiation [or exemplification] is a deeply mysterious relation. If universals are in rebus, then a universal is wholly present in each of its distinct instances. What this could mean is hard to say. [Heil (2003), p. 12]

Third, the realist approach quickly gets tangled up, as Aristotle saw long ago, in undesirable regressions. The recourse to general entities is meant to provide a general account for predication. How is it that a sentence such as “this shirt is red” can be true? Well, the realist says, this is because the object designated by the subject of such a sentence exemplifies the property designated by the predicate. But then, one might ask, how can the sentence “this shirt exemplifies redness” be true? If the original explanation was any good, and if it is as general as it should be, what we should say is that this is because the object designated by the subject of the sentence exemplifies the property of exemplifying redness. And how, in turn, can “this shirt exemplifies the exemplification of redness” be true? This should be because the shirt exemplifies the property of exemplifying the exemplification of redness. And so on ad infinitum. The alleged foundation, then, perpetually slips away.

These are only indicative remarks. Thousands of pages have been written for and against the arguments I have just sketched. But I find these arguments sufficiently convincing to yield the conclusion that the nominalistic approach to thought and language still deserves to be systematically explored. In the last half century, as a matter of fact, nominalism, under one guise or another, has been ably defended and expanded upon by such outstanding philosophers as Nelson Goodman, Wilfrid Sellars, Arthur Prior, Ian
Hacking, and numerous others. It is true, admittedly, that many philosophers nowadays believe nominalism to have been refuted, especially in the philosophy of mathematics and science; one influential argument to this effect is that our mathematics, as it is used in science, presupposes the existence of certain non-singular entities, sets in particular, and that we cannot, therefore, reject this ontological commitment without renouncing science as a whole [see, in particular, Quine (1957) and Putnam (1971)]. Another similarly inspired argument contends that the general adequacy of science presupposes that there truly are laws of nature that apply out there in the world, and that the very idea of a law of nature involves the reality of some sort of universals. [The argument is developed, for example, in Armstrong (1983) and Tooley (1987)]. Any discussion of these arguments, of course, quickly becomes complicated, and I cannot engage in such discussions here. Let me simply mention that, in the last couple of decades or so, there has been an interesting array of theoretical attempts to account for the success of mathematics and general laws in empirical sciences without accepting the ontological commitments to numbers, sets, general properties, and so on that are often supposed to go along with them [e.g., Field (1980, 1989), Heil (2003), Van Fraassen (1980, 1989), Hellmann (1989), Panaccio (1992), especially Ch. 4; Rodriguez-Pereyra (2002)], and that the debate is far from being settled. Whatever the final conclusion turns out to be – if a final conclusion can ever be reached in such matters – it is clear at any rate that all these discussions and explorations of nominalistic ideas about mathematics and science are fruitful and enlightening. And this, I am convinced, can equally be true in the field of cognition theory. What is at stake here is the general intelligibility of our theoretical explanations. Even if the nominalistic outlook should turn out, in the end, to be impracticable – which I strongly doubt – systematic discussion and exploration of this approach is bound to be fruitful and enlightening for the whole domain of cognitive sciences, where realism about properties or attributes is often uncritically – and inadvertently – assumed.

4. Nominalistic constraints for the theory of concepts

In concept theory, the nominalistic option can be cashed out in terms of two main constraints, the distinction between them corresponding to the crucial distinction drawn by Ockham’s Cleaver. The first constraint is that anything which is supposed to occur as a represented thing – i.e., anything which is supposed to be represented by a concept in any way – should be a singular thing. And the second constraint is that anything which is supposed to occur as a representation should also be a singular thing. Let us take a closer look at each one of these in turn.

4.1. Represented things as singular

The first constraint has to do with whatever is represented by a concept. As we have seen, many theoreticians in the field uncritically assume that a concept typically represents a property, or an attribute, or a set, or even a set of properties, and so on. What the nominalistic approach recommends instead, for the sake of intelligibility, is that a concept, insofar as it is taken to be general, must be said to simultaneously represent several things, all of them singular. As William of Ockham wrote long ago, “nothing is general except in virtue of being a sign of several things” [William of Ockham (1974) I, 14, p. 78]. Generality, from the nominalistic point of view, is not an ontological feature, but a semantic – or representational – feature: “The name ‘man’”, Ockham says, “signifies indifferently all particular men” [William of Ockham (1974) I, 17, p. 86]; and the same is true, he adds, of the concept of man. The sign relation, or representational relation, which is needed here is a one-to-many relation. It is what has sometimes been called in twentieth-century semantics multiple denotation or plural reference. What is renounced in such a view is the principle Rudolf Carnap has called the Principle of Univocality, according to which any sign, or name, or representation, should represent exactly one thing [Carnap (1956), p. 98]. Any one of our concepts, instead, can be taken to represent several things at once.

This is not to say that all general concepts must be arbitrary. Accepting such an unfortunate consequence would simply mean renouncing science altogether. But this is by no means called for by the nominalistic outlook. The consequence would follow only on the assumption that in order for any two things to non-arbitrarily fall under the same concept, they must really share something. But this is precisely what the nominalist rejects. Two ontologically distinct things, in this view, can correctly be said to fall under the same concept solely by virtue of resembling each other. Following Bertrand Russell, some authors have objected to this because resemblance, at least, would then have to be admitted as a real universal [Russell (1912), Ch. 10]. But this is not so. Nominalism, as a general doctrine about the basic furniture of the world, is entirely compatible with the idea that all those singular things out there objectively stand in certain relations to each other, such as spatial and temporal relations, causal relations, or resemblance relations. Its point is that this can be admitted without reifying the relations themselves as being additional general things. And this is something the nominalist claims to be able to do: a good semantics can be devised, he contends, for relational statements of the form “x resembles y” without being committed to the real existence of a special single general denotatum for the verb ‘to resemble’.

Another important point to notice is that a nominalistic semantics based on plural reference does not have to be purely extensionalist, in the sense that all representational import for a given concept should reduce to its extension. Purely extensionalist accounts of the semantic import of general terms run into trouble, as is well known,

---

7 For William of Ockham’s theory of relational statements, for example, see Beretta (1999) or Panaccio (2004).
over the cases of coextensive but non-synonymous terms or representations. Suppose, as in Quine’s famous example, that every animal with a heart also has a kidney and *vice versa*. The concept of a [CORDATE] (an animal with a heart) would then have the same extension as the concept of a [RENATE] (an animal with a kidney): the two of them would apply to the same individuals. Yet, the concept of [CORDATE] obviously does not *mean* the same thing as the concept of [RENATE]. Similarly, there are no unicorns, and no centaurs either. The concepts of [UNICORN] and [CENTAUR], therefore, have the very same extension, namely the null extension. Yet they are not representationally equivalent. There has to be more to conceptual representation than mere extension. This is why many authors feel that a good semantics, whether for words or for concepts, needs another relation, in addition to being true of, or applying to – a relation they sometimes call *connotation*.

William of Ockham thought so too. Actually, he was probably the first to systematically insist on the need for such a relation – which, in fact, he labeled *connotatio*. But unlike many later authors, he gave it a nominalistic interpretation. Realist philosophers usually think of connotation as a link between a representation and a general property or an abstract object. While the extension of [WHITE], they would say, contains every singular white thing, what the term connotes is the general property, or shared attribute, of whiteness. This is not how Ockham conceived of connotation. The medieval philosopher, once more, has an intriguing suggestion for us. Connotation, he says, just like signification or reference, normally connects a sign or representation with *several singular things* at once. Take [WHITE] again, which Ockham considers as a good example of a connotative term. What it primarily represents – or signifies – are all the white things there are, each one of which, of course, is fully singular. But in addition to primarily signifying white things, Ockham says that [WHITE] connotes their individual whitenesses, with each white thing being taken to have its own singular whiteness. Or take [RENATE] and [CORDATE]. They both refer – *ex hypothesi* – to the same things. Where their representational import differs is in their connotation: [RENATE] connotes kidneys, and [CORDATE] connotes hearts. But kidneys and hearts are singular things; they are no more intrinsically general than the animals that have them. The distinctive feature of connotative concepts, according to Ockham, is that they primarily represent certain singular things – those things that they apply to, or are true of – by associating each one of them with some other singular thing(s): a particular whiteness in the case of the concept [WHITE], a kidney in the case of [RENATE], or a heart in that of [CORDATE].

How this account can be extended to relational concepts is straightforward. A concept like [FATHER OF] – another of Ockham’s favorite examples – will be said to primarily signify every father and to connote every child, in such a way that it associates each one of its primary significates with at least one of its connotata. The details have to be worked out, of course. But the main point, by now, should be clear. Applying the

---

Ockhamistic idea of connotation to the realm of conceptual representation provides a semantic tool for understanding how a concept can simultaneously represent several pairs – or trios or quartets, for that matter – of ordered individuals (pairs of white things with their whitenesses, for example, or pairs of cordate animals with their hearts, and so on) without countenancing any real general property, attribute, or relation. If it is any good at all, this sort of explanation should hold in particular for the crucial relational concept of [RESEMBLANCE] itself and help us see how statements of the form “x resembles y”, which are so important for the nominalist, could turn out to be true without a general relation of resemblance having to be admitted as an extra real thing out there in the world. This whole nominalistic idea of connotation, in short, stands out as an intriguing idea, which deserves to be explored and worked out, even though it has unfortunately been neglected in recent semantics. Its great merit for cognitive science is that it gives a clue as to how a semantic account of mental representation might be devised with only singular beings as represented things.

4.2. Representations as singular

The second main constraint on a nominalistic theory of concepts is that representations themselves should not be taken as anything but singular things. A concept, in this view, is always located within a given singular mind. It is a singular state of a given thinking agent. Your concepts are states of your own mind, or – plausibly enough – of your own organism, and my concepts are states of my mind, or organism—neurological states, presumably. In a sense, then, according to this nominalistic perspective, your concepts are irreducibly distinct from mine, and vice versa. This is not to say that they cannot be semantically equivalent. Nothing in the nominalistic view prevents your concepts from being representationally equivalent to mine. One of yours could represent and connote exactly the same individual things as one of mine, and in the same ways. Yet ontologically, they will be distinct singular things, exactly as two tokens of a word (cat and cat, for example) are two distinct utterances or inscriptions even though they can be treated as entirely equivalent to each other from a semantic point of view. Using the terminology of tokens and types, what nominalism claims is that, among concepts as among words, only singular tokens exist.

This, admittedly, raises the delicate question of how distinct tokens are grouped into types. If my concept of [DOG] is ontologically distinct from your concept of [DOG], how can we say that, in a sense, we both have the same concept? This is certainly something we should be able to say in a sense, if we are interested in a theory of concepts at all. And it is actually the kind of thing that researchers in the field say all the time. They speak of the concept of [DOG], how people acquire it, or how they use it (in the singular). How is that possible and acceptable if only tokens really exist? This challenge has to be answered, since we cannot simply abandon all talk about types as mere nonsense. Another way of putting the same problem is the following. It is often said – by Jerry Fodor, for instance – that “concepts are public; they’re the sorts of things that lots of people can, and do, share” [Fodor (1998), p. 28]. Fodor even counts this as one of the “not-negotiable conditions on
a theory of concepts” [Fodor (1998) 23]. On a nominalistic approach, however, this should be the object of some negotiation. If concepts are mental states, they are not shared, strictly speaking: you have your states, and I have mine. The real problem is: under what conditions can they be said to be representationally equivalent?

To get a hint of how such a question can be dealt with, let us turn for a moment to a theory which, apparently, does not have this sort of problem, namely a Fregean theory of concepts as mind-independent abstract objects. Concepts, in this view, are literally public, in the sense that they can be grasped by different people, simultaneously or at different times. In such an ontology, the concept you grasp and the concept I grasp can literally be the same: when we both grasp the concept of [DOG], we are both aware of the same abstract entity. But let us now take our reflection one step further. Even in this Fregean picture, your grasping of the concept of [DOG] and my grasping of the same concept are distinct states: your grasping is a state of your mind, and my grasping is a state of my mind. What makes them equivalent from a semantic point of view is that their extramental object is the same. This is OK. But why, then, should a similar idea not be exploited in a nominalistic vein? Think of concepts, now, as being mental states themselves rather than objects of mental states. Why is it that a certain mental state of yours can be said to be a concept of [DOG]? Well, in a nominalistic spirit, that is because this particular state of yours represents dogs somehow. And why is it that a certain state of mine can be said to be my concept of [DOG]? It is because it represents dogs too. So what makes your concept and mine equivalent in this view is that they represent the same singular things. And if you want to add connotation to this picture, you are welcome to do so. Our concepts of a white thing, let us say, are equivalent, although ontologically distinct, because they both apply to the same singular things out there in the world – namely the white things – and they both connote the same singular things as well – namely the singular white things. The principle here is the same as in the Fregean approach: our distinct mental states are taken to be semantically equivalent because their external objects are identical. Note, however, that the nominalistic explanation seems to be both more economical and more enlightening than the Fregean one, since it dispenses with intermediate abstract objects such as the abstract concept of [DOG], and also with the alleged states of grasping such abstract objects. Note also that the publicness of concepts in this approach has finally been negotiated a bit since what turns out to be public are the represented things, not the representing states themselves.

It must be added, though, that very often in the literature and in the experiments conducted by concept theorists, concepts are identified – or grouped into types – not so much on the basis of what their represented objects are, but on the basis of what public words they are associated with. Authors will often speak of somebody’s concept of [DOG], for example, as the mental representation this person associates with the word dog (or with certain uses of the word dog). In a sense, this is very different from the situation in the previous case where the public identity of the concept type depended upon the represented objects. And this difference must be insisted on because it often goes unnoticed in the literature – which turns out to be quite harmful. On the other hand, there is an important similarity between the two cases which must also be stressed. In
the new case, as in the previous one, the grouping of ontologically distinct mental representations into one single type depends upon the independent identifiability of external and publicly observable objects; these objects, in the new case, are words rather than the represented singular things. This difference, of course, complicates the situation a bit. Not only do we now have two non-equivalent ways of grouping concepts into types (on the basis of their representational import, on the one hand, and on the basis of what public word they are associated with, on the other hand), but in the new case, in addition, the identifying criterion – namely the external word – also presents the character of a type, the concept of [DOG] being identified as such with reference to the word *dog*. But the word *dog*, in turn, is not something that enjoys any independent existence, according to the nominalist. The typing problem, then, reappears under a new guise: how is it that different spoken or written tokens can legitimately be grouped into linguistic types? And the answer, of course, cannot simply be that it is because they express the *same* concept, for we would then be engaged in a vicious circularity. So we do have a complication here. But it is a complication of the situation itself. That nominalism helps to identify it is a virtue rather than a defect. Speaking uncritically of the property of doghood, *the* concept of [DOG], or *the* word *dog* masks a bunch of genuine problems for the theory of concepts and favors confusion. Whether one ends up entirely endorsing the second nominalistic constraint – concerning the basic singularity of mental representations themselves – or not, it should be useful and enlightening to at least think about it, and to try to enforce it as much as possible.

Generally speaking, nominalistic concerns with respect to concept theory have to do with exactly what sorts of things we think we are talking about when we speak of concepts and what they represent. Keeping such concerns in mind helps us to develop a healthy awareness of certain important problems and distinctions, and avoid certain widespread confusions. This, in the end, is the point I have tried to make here, by explaining how a nominalistic outlook can bear upon the way we think about concepts as mental representations.

References

Beretta, B. (1999), Ad aliquid. La relation chez Guillaume d’Occam (Éditions Universitaires, Fribourg, Switzerland).


Panaccio, C. (1992), Les mots, les concepts et les choses. La sémantique de Guillaume d’Occam et le nominalisme d’aujourd’hui (Bellarmin/Vrin, Montréal/Paris).


William of Ockham (1991), Quodlibetal Questions [Translated by A.J. Freddoso and F.E. Kelley, 2 volumes (Yale University Press, New Haven, CT)].
Chapter 46

WHY DO WE THINK Racially?

EDOUARD MACHERY
University of Pittsburgh, History and Philosophy of Science

LUC FAUCHER
Université du Québec à Montréal, Philosophy

Contents

Abstract 1010
1. Introduction 1011
2. Is racialism a mere social construct? 1012
   2.1. Racial skepticism 1012
   2.2. Races are interactive kinds 1013
   2.3. Races are transient kinds 1013
   2.4. Merits and problems 1015
3. Is racialism a by-product of a human kind module? 1016
   3.1. The nature of racialism 1016
   3.2. The human kind module 1017
   3.3. Empirical evidence 1017
   3.4. Merits and problems 1019
4. Are races mere coalitions? 1021
   4.1. Races and coalitions 1021
   4.2. Empirical evidence 1021
   4.3. Merits and problems 1022
5. Is racialism a by-product of an evolved ethnic cognitive system? 1024
   5.1. “Ethnies” are not mere coalitions 1024
   5.2. An adaptive scenario: Ethnic cognition and the exaptation of human folk biology 1025
   5.3. Empirical evidence 1026
   5.4. Merits and problems 1027
6. Conclusion 1029
References 1031
Abstract

Contemporary research on racial categorization is mostly encompassed by two research traditions – the social constructionist approach and the cognitive-cum-evolutionary approach. Although the literature from both approaches has some plausible empirical evidence and some theoretical insights to contribute to a full understanding of racial categorization, there has been little contact between their proponents. In order to foster such contacts, we critically review both traditions, focusing particularly on the recent evolutionary/cognitive explanations of racial categorization. On the basis of this critical survey, we put forward a list of eleven requirements that a satisfactory theory of racial categorization should satisfy. We conclude that despite some decisive progress, we are still far from having a complete theory of why humans classify people on the basis of their skin color, body appearance, or hair style.
1. Introduction

One of the most important challenges facing the psychology of categorization is to explain why, of all the ways of dividing the world, we find that some categories, but not others, are natural. Psychologists usually underscore two kinds of causes: first, the nature of human cognition, and second, the nature of the world itself [Malt (1995)]. Social and cultural factors are usually thought of as secondary factors, if they are mentioned at all.

The study of social categories, for example, the category of teenagers [Hacking (1999)], has shown the shortcomings of this approach. It is difficult to account for these categories without paying attention to the social circumstances that surrounded their formation. To understand social categories, and possibly for to understand other categories too, social, cultural, and psychological factors all have to be taken into consideration.

The study of the concepts of races, for example, [BLACK], [WHITE], etc., and the concept of race itself, that is, how people conceptualize race membership, illustrates the need for a more integrative approach to categorization. As we will see, an integrative approach is required to explain the nature and origins of racialism – that is, the fact that people classify humans on the basis of their visible physical properties (skin color, body shape, height, hair appearance, etc.) and believe that this classification picks out meaningful, important biological kinds. This integration has been hindered by the fact that the study of racialism is scattered over several fields, particularly history, anthropology, sociology, and psychology. Moreover, researchers are often committed to different theoretical tenets. Contemporary research on racialism is mostly encompassed by two research traditions – the social constructionist approach and the cognitive-cum-evolutionary approach. Most social constructionists believe that the concept of race is a pseudo-biological concept that results from specific historical circumstances, and that it has been used to justify the unequal treatment of specific groups of people. Proponents of the cognitive-cum-evolutionary approach, on the other hand, believe that racialism results from a specific cognitive system with a particular evolutionary history. Since the tenets of these two research traditions are often judged to be inconsistent, there has been little contact between the proponents of these two approaches. This is certainly unfortunate, for the literature from both approaches has plausible empirical evidence and theoretical insights to contribute to a full understanding of racialism.

Like others, we believe that the time has come to bridge the gap between these two research traditions [Sperber (1996), Faucher (1999), Fiske (2000), Mallon and Stich (2000), Boyer (2001a), Medin and Atran (2004), Machery and Faucher (in press)]. This chapter is a step in this direction. We review some recent and significant contributions

---

1 We use capital letters within square brackets to indicate concepts.
2 Racialism must be distinguished from racism, which adds value judgments (mostly negative, but sometimes positive) to racialism. In this article, we focus on racialism. Most social scientists focus on racism and take the existence of racial classifications for granted, without inquiring about their origin.
to the contemporary literature on racialism\(^3\). By doing this, we hope to stimulate research projects on racialism that would take into consideration both research traditions. We present the social constructionist approach and we discuss in more detail three cognitive-cum-evolutionary theories. We underscore the empirical and theoretical problems of current theories of racialism. On the basis of this critical survey, we put forward a list of 11 requirements that a satisfactory theory of racialism should satisfy. We conclude that despite some decisive progress, we are still far from having at our disposal a complete, satisfactory theory of why humans classify people on the basis of their skin color, body appearance, and other physical properties.

2. Is racialism a mere social construct?

2.1. Racial skepticism

A dominant view about races nowadays is the so-called “social constructionist view.” Like many scientists, social constructionists argue that the concept of race does not have any biological reality [e.g., Appiah (1996)]. Since the 1970s, it has been widely recognized that the biological concept of subspecies could not be applied to humans. First, there is more genetic variability within human racial groups than between them [Lewontin (1972), Brown and Armelagos (2001)]. Moreover, classifications based on different phenotypic traits (skin color, body shape, hair appearance, etc.) usually cross-cut each other. Finally, these traits do not correlate in a systematic way with other biological characteristics [Diamond (1994), but see also the discussion in Nature Genetics, Supplement, November 2004]. Hence, assigning an individual to a race does not buy the inferential power one is usually warranted to expect from a biological kind term\(^4\).

Social constructionists about race are not mere skeptics. They usually also underscore the instability and diversity of human beings’ concepts of race\(^5\). For instance, Omi and Winant remark that an “effort must be made to understand race as an unstable and ‘decentered’ complex of social meanings constantly being transformed by political struggle” (2002, p. 123; our emphasis). Others suggest that the notion was invented by a specific culture at a specific time, for instance during the late fifteenth century [Fredrickson (2003)]. How should we interpret such statements? A careful examination of social constructionism is in order.

\(^3\) For the sake of space, we do not review the literature in social psychology that treats racial prejudice as a special kind of prejudice [Crandall and Eshleman (2003)] nor the literature that explains racial classification as a by-product of our perceptual categorization system [Aboud (1988)].

\(^4\) If one gets any inferential power at all, this may be mostly because the concept of race “continues to play a fundamental role in structuring and representing the social world” [Omi and Winant (2002), p. 124; see also Root (2000)].

\(^5\) In what follows, the term “race” is used to refer to the groups that are identified as racial by some societies. Although there are no races, there are groups that are identified as races, e.g., Blacks, Whites, etc.
2.2. **Races are interactive kinds**

In this and the following sections, we use some distinctions made by the philosopher Ian Hacking to clarify the social constructionist position about races\(^6\). Once these distinctions are in place, it will be easier to identify the strengths and shortcomings of this approach.

Hacking defends a new brand of nominalism, called “dynamical nominalism.” This nominalism is more selective than its historical predecessors, for it applies only to kinds of human beings, the human kinds\(^7\). As Hacking puts it:

> I hold that strict nominalism is unintelligible for horses and planets. … Gloves are something else: we manufacture them. I know not which came first, the thought or the mitten, but they evolved hand in hand. … My claim about making up people is that in a few interesting respects multiple personalities (and much else) are more like gloves than like horses.\(^{1986, p. 229}\)

This claim draws on Hacking’s distinction between “indifferent kinds” and “interactive kinds.” Members of indifferent kinds, e.g., rocks, are not affected by the way we categorize them. On the contrary, we, human beings, react to the categories that are used to classify us. We care about and can be transformed by the way we are categorized. We are responsive to labels, concepts, and beliefs about the kinds that are applied to us. In Hacking’s words, we are “an interactive kind” (1999, p. 34).

For Hacking, races are among the paradigmatic human kinds [see also Root (2000)]. For, as we saw, there is no fact of the matter that would define races independently of what we believe about them. Thus, the properties that characterize each race are created together with the concept that refers to it.

Hacking’s notion of interactive kind captures an important aspect of the social constructionist view of races. For social constructionists, when the concept of race is applied to us, it changes the way we think about ourselves and, as a result, what we do, what we learn, and what we are. Thus, racial identities are constantly shaped by our concepts of, and our beliefs, about races. Omni and Winant call this phenomenon “racial subjection” [see also Holt (1995), Appiah (1996), Mallon (2003)].

2.3. **Races are transient kinds**

Hacking also draws a distinction between “transient” and “permanent” kinds. A kind is transient if its reality depends on a cultural niche – that is, on the ideas and social practices that are prevalent in a given society at a given time. The dependence is such that were the niche of a transient kind to disappear, the kind would disappear as well.

---

\(^6\) Hacking (1999) rejects the label “social constructionism.” We believe, however, that Hacking captures the core insights of the social constructionist approach.

\(^7\) Hacking (1995) uses “human kind” to refer to the kinds picked out by social sciences. We use this expression in a more liberal way to designate any category intended to group humans. There is no reason to restrict the effects described by Hacking to the kinds used in the social sciences.
Hacking suggests, for instance, that this happened with the concept of hysteria at the beginning of the twentieth century. In contrast, the reality of permanent kinds—basically, natural kinds like tigers—is not dependent on any culture, even if the concepts that refer to them may vary across cultures [Hacking (1999)].

Human kinds typically depend on culture- and time-specific beliefs. Thus, they are typically transient kinds [Hacking (1995)]. The notion of transient kind is useful to clarify an important aspect of the social constructionist view of race. Most scientists, social constructionist or not, recognize that human races are not natural kinds. In addition, social constructionists believe that races should be conceived as transient kinds. There are different ways of “cashing out” this idea, some more radical than others.

Many constructionists insist on the cultural and historical differences between culture- and time-specific conceptualizations of race membership. Like other concepts of transient kinds, the way race is conceptualized—and, as a result, race itself—depends on specific cultural niches. For example, Pascoe (1996) argues that in the United States, a dichotomous conceptualization of race membership became prevalent after the abolition of slavery. In that view, race membership could not be mixed or graded. The “One drop of blood” Rule, according to which a person was black if he or she had one black ancestor, and the infamous miscegenation laws that forbade interracial sex and marriage, illustrate this conceptualization. Support for this conceptualization weakened in the 1950s and 1960s. Sociologists have also shown that nowadays, more and more Americans use mixed racial categories to describe their racial membership [Skidmore (1993), Aspinall (2003)]. They propose that this evolution is related to new forms of immigration and to various social practices [e.g., writings emphasizing mixed racial identities; see Aspinall (2003)]. In addition, many studies illustrate the differences between culture-specific conceptualizations of race membership. A classic example is the difference between Brazilian and American concepts of race [Harris (1963), see the nuanced review in Skidmore (1993) and the critical discussion of Gil-White (2001b)].

We maintain that these social constructionist views are moderate, for they claim only that the concept of race goes through many different elaborations. This is compatible with the idea that there is something common to these different elaborations. These commonalities could result from some aspect of human cognition. Moderate social constructionism is thus consistent with the idea that because of some psychological disposition, humans tend to classify people into races when they meet other people with different phenotypes.

Some constructionists go further. They believe that the concept of race is a recent, Western invention. Many of them argue that there were no concepts such as [WHITE PERSON] or [BLACK PERSON] until theories of race appeared in Europe in the eighteenth century [for a different origin, see Fredrickson’s (2003) article]. Greeks, for instance, might have divided people into some categories, but they relied on nonracial categories. We are told that they were indifferent to skin colors; instead, they were more preoccupied by geographic origins[^8].

Some scientists have, for instance, argued that racialism results from the application of the principles of scientific biological classification to humans. During the eighteenth century, scientists like Linnaeus and Blumenbach developed the principles of biological classification. These principles were applied by scientists – biologists, geographers and, in the nineteenth century, anthropologists – to humans on the basis of their phenotypic properties [Gould (1994)]. This is known as “scientific racialism.” Scientific racialism became part of the European culture as a justification for the colonization of the rest of the world\(^9\).

We maintain that this kind of constructionism is “radical,” since it affirms that the concept of race is exclusively the product of historical and cultural causes. It claims that humans do not tend to classify people into races when groups with different phenotypes meet, save for particular historical circumstances.

### 2.4. Merits and problems

The social constructionist approach is obviously important. Constructionists have correctly insisted on the cultural aspect of humans’ concepts of race and on how racialist classifications affect people. Some constructionists have illustrated the diversity of humans’ concepts of race. It is obviously important to distinguish between different time- and culture-specific conceptualizations of race, if only because they determine what kind of identity is available for individuals at a given time and place. Finally, they have shown that racialist distinctions have been used for social and political purposes.

The radical version of social constructionism is, however, extremely problematic. According to this form of constructionism, it should not be the case that many non-related cultures at different times have developed some racial classification. Hence, if something like the concept of race appears in many nonrelated cultures, the radical thesis is falsified. Now, contacts between groups that are characterized by different phenotypes are plausibly a relatively recent phenomenon. There is some historical evidence, however, that when these contacts happen, humans tend to classify racially. For instance, around the twelfth century, when contacts with foreigners first became common in China, the color-consciousness of the Chinese people increased strongly, and the white skin of Occidentals – which was called “ash-white” skin – was thought of as “the exteriorization of the demonological forces” [Dikötter (1992)]. Dikötter also notes that the Chinese had equated [BLACK] and [SLAVE] at a quite early stage of their history. Moreover, Isaac (2004) has recently provided some convincing evidence that racialism and racism also existed in classical Greece and Rome.

Moderate social constructionism has a different kind of problem. It has focused its attention on differences between cultures or times, but it has left unexplained the commonalities between culture-specific concepts of race [Machery and Faucher (in press)].

---

\(^9\) Although we do not deny that the European eighteenth- and nineteen-century conceptualization of race membership is specific, as we go on to explain, we reject the idea that racialism is a mere invention of Europeans in the nineteenth century.
In fact the content of human beings’ folk concepts of race seems to be tightly constricted in some respects [Hirschfeld (1996)]. Particularly, the social constructionist approach does not explain why races are cross-culturally thought of as biological kinds. Finally, moderate social constructionists illustrate the diversity between concepts of race, but fail to explain it.

There is a growing body of literature in evolutionary psychology and evolutionary anthropology that bears on these issues. Although no consensus has yet emerged, several recent proposals attempt to describe the cognitive mechanisms that produce the concept of race [Hirschfeld (1996), Kurzban, Tooby and Cosmides (2001), Gil-White (2001a), Cosmides, Tooby and Kurzban (2003)].

3. Is racialism a by-product of a human kind module?

Before looking at the differences between and limits of these evolutionary and cognitive hypotheses, we want to emphasize what they have in common. They all endorse the following four claims:

● Racial skepticism: there are no races.
● Racialism does not result from a domain-general, perceptual categorization system.
● Racialism results from a domain-specific cognitive system. However, cognitive and evolutionary-minded scientists disagree on the nature of this system.
● Racialism has not been selected for: it is a by-product of a cognitive system that has been selected for other reasons.

We turn now to Lawrence Hirschfeld’s research program.

3.1. The nature of racialism

According to Hirschfeld (1996, 1997, 2001), young children, across cultures, believe that races are characterized not only by bundles of race-typical properties – e.g., phenotypic properties like skin color and body appearance – but also by bundles of behavioral and psychological properties. Moreover, young children essentialize races: they assume that races are defined by immutable, natural, and often unknown essences that are passed down from parents to children [Hirschfeld (1996)]. These essences explain why people possess the assumed race-typical properties. As a result, these race-typical properties are believed to be inherited and to be insensitive to people’s rearing environment. Hirschfeld notices that this is in some aspects similar to the way children think about animal species, although he argues that racial cognition is not derived from our biological cognition.

Essentialism fosters inductive generalizations [Hirschfeld (1996)]. People expect members of a given race to have many properties in common. As a result, they draw inductions from a limited number of instances. Properties that are true of most members of a race, say, poverty, can also be explained in racialist terms – as the natural effect of a hidden essence.
Children manifest this view of race early in their lives [Hirschfeld (1996)]. It does not result from their parents’ explicit teachings. Most parents do not teach explicitly racialist distinctions. When they do, young children’s racialist attitudes tend not to reflect those of their parents [Hirschfeld (2001), but see Crandall and Eshleman (2003)]. Moreover, children do not acquire their concepts of races from people’s physical appearances [Hirschfeld (1996)]. Thus, their concepts are not derived from the domain-general, perceptual processing of categories. On the contrary, young children seem to rely on linguistic cues to form concepts of races. Hirschfeld concludes that racialism results from an innate, domain-specific and universal cognitive system.

3.2. The human kind module

What is this cognitive system? According to Hirschfeld, human beings have evolved a folk sociology, also called the “human kind module” [Hirschfeld (1996, 2001), Sperber and Hirschfeld (2004)]. Its proper domain consists of the social groups that constitute a society. It includes kin-based groups and coalitions. Recently, Hirschfeld (2001) has emphasized coalitions among the proper stimuli of this module. He claims that coalitions are much more important in the human species than in any other animal species. Humans belong to an astonishing number of coalitions that cross-cut each other and very often shift. This has created some evolutionary pressures for developing a cognitive system dedicated to track coalitions [Hirschfeld (1997, 2001)].

This human kind module does not determine which groups – particularly, which coalitions – are relevant in a society. It interacts with specific cultural environments, in which some specific groups are salient. These culture-specific groups are essentialized [Hirschfeld (1997)].

Thus, we are not innately predisposed to adopt racialist classifications. In some societies, however, coalitions are formed on the basis of superficial phenotypic properties – for example, skin color. In those societies, children pick out the groups that are formed on this basis – i.e., races – and conceptualize them in an essentialist manner. If such groups were not coalitions in our societies, children would not draw racialist distinctions. Instead, they would pick out other coalitional groups. For example, Hirschfeld (1997, 2001) reports that children in India essentialize and naturalize caste and occupation.

3.3. Empirical evidence

Hirschfeld’s theory is supported by several experiments [see Hirschfeld (1996)]. We describe the two most important ones.

---

10 That is, the stimuli that the module has evolved to deal with. The proper domain of a module contrasts with its actual domain, that is, the stimuli that it deals with now, even if it has not evolved to do so [Sperber (1996), Sperber and Hirschfeld (2004)].

11 Hirschfeld (1996) also emphasizes the results of the “switched-at-birth” experiment. Since Gil-White’s work is based on this experiment, we will present its structure in Section 5.
In the first experiment (1996), 3- and 4-year-old (and slightly older) children in the United States and in France were presented with three types of drawings that depicted an adult with two children. In the first drawing, one child had the same skin color (black or white) as the adult, but had a different body build (thin or fat), while the other child had the same body build as the adult, but had a different skin color. In the second drawing, one child had the same skin color as the adult, while the other child wore the same professional uniform (e.g., a police uniform) as the adult. In the third drawing, one child had the same body build as the adult, while the other child wore the same uniform as the adult. Thus, three properties were being contrasted: a racial characteristic, a physical characteristic, and a profession. Some children were asked question (1): “In the drawing, which of the two children represented the adult as a child?” Other children were asked question (2): “In the drawing, which of the two children was the adult’s child?”

In response to question (1), race (skin color) was chosen over occupation and body build as a cue for identity across times. Although younger children were less likely than older children to choose race over occupation and body build, 3-year-olds chose race over body build at a rate that was significantly above chance, and 4-year-olds chose race over occupation at a rate that was significantly above chance. The same pattern was also found for responses to question (2).

Hirschfeld concluded that even young preschoolers believe that a racial characteristic (skin color) is constant during growth and transmitted by descent [but see Solomon et al. (1996)]. This is not the case for other properties, like body build and profession [Hirschfeld (2001)]. Since this kind of reasoning is assumed to be build on a belief in essences [Gelman and Wellman (1991)], Hirschfeld infers that children essentialize race membership, that is, they believe that races are defined by essences transmitted by descent that cause the possession of the racial phenotypes.

The second experiment concerns the origin of racialist classifications [Hirschfeld (1996)]. In the first part of the experiment, children were told a story about a child who had to buy a scarf for his mother. The child asked four adults for help. Each adult was described twice in terms of behavior, occupation, clothes, gender, race, and body type. Children were given a free recall test immediately afterward. The point was to see what kind of information had been encoded. The results showed that even 3-year-old children encoded each of the dimensions (behavior, occupation, clothes, gender, race, and body type). Not all types of information, however, were equally encoded. Four-year-old children encoded occupation more than race, and race more than gender and body type.

These results were later compared with those of the second part of the experiment. In the second part of the experiment, children were shown a set of drawings that told a story that was similar to the story used in the first part of the experiment. They were asked to describe the events portrayed by the drawings. Their memories of the story were also probed. The results showed the following. When they described the drawings, 3- and 4-year-old children mentioned the characters’ gender more than their profession. Race was rarely mentioned. In the recall task, children mostly mentioned the characters’ behavior, their gender, and their clothes. Again, race was very rarely mentioned.
Thus, both tasks revealed very few references to the characters’ race. There were no significant difference in responses between 3- and 4-year-olds.

From both parts of the second experiment, Hirschfeld concluded that external or physical appearance is not a crucial component of children’s initial racialist concepts. [WHITE], [BLACK], [CHICANO], etc., are not perceptual concepts. For, if they were, race membership would be visually salient. Race would be mentioned in a visual description task and in a visual memory task. This was not the case. On the contrary, race was a salient property in a verbal task.

Hirschfeld (1996) inferred that concepts of races are not derived from the perceptual categorization system. Otherwise, the perceptual information would be encoded. Hirschfeld took this to imply that racialist concepts and our conceptualization of race membership result from a domain-specific cognitive system that is preferentially triggered by verbal cues – the human kind module.

3.4. Merits and problems

We view Hirschfeld’s research as an important contribution to the understanding of racialism. He has provided a large set of data that have to be taken into account by any theory of racial cognition. He has also put forward some important theoretical points. Let us quickly summarize the main findings that have to be kept in mind when discussing a theory of racial cognition:

(1) In several respects, people reason about races as they reason about animal species.
(2) There is some evidence that children manifest this bias at a young age (as young as 3 years old, for some tasks).
(3) Children’s racialist cognition is not derived from the domain-general categorization system that processes visual inputs. Linguistic inputs may be crucial.
(4) Points 1 to 3 suggest that racialism results from some domain-specific cognitive system.

Despite its importance, Hirschfeld’s approach is nonetheless problematic. We focus first on some theoretical problems and then turn to look at some empirical ones.

First, in Hirschfeld’s approach, the hypothesized human kind module is underspecified. It is unclear why our social cognition is similar in certain aspects to our biological cognition. Moreover, it is unclear why not all social groups are thought of biologically, if racialism is a by-product of our folk sociology. Hirschfeld (1996) claims that this human kind module picks out social groups that are salient in a given society. However, there are plenty of salient social groups – e.g., neighborhoods – that are not thought of in the same way as animal species are. To counter this objection, Hirschfeld should say that this human kind module picks out a specific kind of group. In more recent articles, Hirschfeld has indeed suggested that the function of this module is to track coalitions [Hirschfeld (2001)]. However, as we shall see in Section 4 discussion of a similar proposal by Kurzban et al. (2001), this is not satisfactory.

We will conclude that racialism probably results from another kind of cognitive system.
Second, Hirschfeld does not explain why race is conceptualized differently across cultures. He pays lip service to the project of integrating the social constructionist tradition with the cognitive-cum-evolutionary approach. In fact, however, he does little to really integrate the two perspectives.

Third, Hirschfeld’s proposal is committed to the theory of psychological essentialism [Medin and Ortony (1989), Gelman (2003)]. According to this theory, people believe that many categories are characterized by an essence. Such a folk belief is supposed to explain several reasoning patterns, including category-based induction [Gelman and Markman (1986)] and a commitment to the transmission of properties by descent [Gelman and Wellman (1991)]. We are skeptical about this. We doubt that most people believe in some inner, maybe unknown, property that defines the identity of categories and explains the possession of observables properties. People do not have to believe in essences in order to display the reasoning patterns mentioned above [Strevens (2000, 2001), see the reply by Ahn et al. (2001)]. Thus, Hirschfeld’s data show that races are thought of in the same way that animal species are. But this does not entail that people are committed to essentialism.

We are also puzzled by some aspects of Hirschfeld's experiments. Consider particularly the first experiment described above. Hirschfeld asked children whether a profession, which was symbolized by a professional outfit, was more likely than skin color or body build to be transmitted by descent and to be constant over life. We wonder how these children interpreted this question (for an intriguing comment on their answers, see [Gil-White (2001a)]. Could these children really have believed that a child has a profession?

Second, we are astonished by the discrepancy between Hirschfeld’s ambitious theory and his empirical evidence. More predictions should be derived from his theory and they should be submitted to more stringent tests involving various experimental paradigms. For example, Solomon et al. (1996) have provided some empirical evidence that before age seven, children may not understand that skin color is inherited from one’s biological parents. In their review, Crandall and Eshleman (2003) mention some research that is prima facie inconsistent with Hirschfeld’s claim that children’s racialism is not affected by explicit parental teaching. Thus, it seems too early to be fully convinced by Hirschfeld’s empirical evidence. Moreover, Hirschfeld claims that the hypothesized human kind module is universal. However, he provides little cross-cultural evidence to support this claim, relying instead on children from two industrial societies, France and the United States.

Third, in most experiments, not all children gave the same answers. Consider, for example, the first experiment reported in Section 3.2. Hirschfeld underscored the fact that around 80% of the 7-year-old children believed that a racial property (skin color) was more likely than profession to be constant over life. Thus, 20% of these children gave the opposite answer. Hirschfeld did not say much about the minority answer, treating it in fact as noise. We feel that this approach is mistaken. Individual differences may provide some crucial information about the cognitive system that underlies racialism. Particularly, it could cast some light on the kind of inputs that affect its development – for example, on the importance and nature of the cultural inputs. Moreover, it is a crucial piece of evidence
for evaluating claims about the innateness and evolved nature of cognitive systems. To be fair, though, this way of proceeding is common in many developmental studies.

Hirschfeld’s account has other problems. To begin with, Hirschfeld does not explain why races are so often salient groups. Besides, we have strong doubts concerning his account of the evolution of racial cognition. Since the account by Kurzban et al. (2001) suffers from the same problems, we will discuss this in the next section.

To conclude, we reject the claim that racialism results from the interaction between an essentialist folk sociology and the social structure of some societies. We also emphasize that more evidence is needed to fully support Hirschfeld’s other points.

4. Are races mere coalitions?

4.1. Races and coalitions

In two recent papers [Kurzban et al. (2001), Cosmides et al. (2003)], Kurzban et al. proposed an original hypothesis. Like Hirschfeld, they recognize that it is implausible to posit a race module. First they argue, if there are no races, how could a mechanism that specializes in tracking them have evolved? Second, given that during the evolution of our species, our ancestors had a very meager chance of meeting people with a different skin color, “there could have been no selection for cognitive adaptations designed to preferentially encode such a dimension” [Kurzban et al. (2001), p. 15387].

Instead, Kurzban et al. propose that racial categorization is a by-product of a cognitive system that evolved to detect coalitions and alliances. Since in multiracial, non-integrated societies, races are coalitions, they are picked out by the coalitional mechanism. According to Kurzban et al., this mechanism has to be sensitive to patterns of cooperation and competition, as well as to cues that predict group allegiances. The mechanism ought to pick up any observable feature (dress, badges, manner, dialect, skin color, etc.) that indicates group allegiance. Moreover, since coalitions are not necessarily stable, we should be able to dynamically revise any coalitional identification.

On the basis of this hypothesis, Kurzban et al. (2001, p. 15389) predict that a “new social environment in which coalition is uncorrelated with race should weaken the pre-existing weight given to race”.

4.2. Empirical evidence

To test the prediction that a change in social environment would weaken the weight given to race as a cue to coalition, Kurzban et al. (2001) used the memory confusion protocol. This test is used to reveal which categories are used to categorize individuals. In the first part of the test, subjects are presented with sentences that are paired with pictures of the individuals who uttered them. Subjects are asked to form an impression of these speakers. In the second part of the test, subjects are given a surprise recall task: they have to recall who said what. It is believed that misattribution reveals encoding. Subjects confuse
the individuals that are classified in the same category more than the individuals that are sorted in different categories. For instance, if the subjects have classified individuals by gender during the first part of the task, they more frequently misattribute a sentence that a man said to another man compared with a sentence that a man said to a woman. The errors should be random for the categories that are not encoded. Previous experiments have shown that race is spontaneously encoded by subjects, and efforts to reduce this spontaneous encoding have been largely unsuccessful.

In Kurzban et al.’s first experiment, experiment (1), subjects were presented with eight sentences paired with the pictures of eight speakers. The pictures represented young men dressed identically. From the sentences, one could infer that they divided into two groups that had been involved in some kind of fight. Their “races” (White and Afro-American) could be seen on the pictures. Subjects were given a surprise recall task, with the pictures still displayed in front of them. Kurzban et al. wanted to know whether subjects could encode the coalition membership solely on the basis of the content of the utterances (no visible cues were correlated with coalition membership). Experiment (2) was identical to experiment (1), except that Kurzban et al. used an arbitrary and nonpermanent marker (the color of a tee-shirt) as a cue of allegiance. If races are mere coalitions, providing relevant coalitional information should decrease the reliance on race as a proxy for coalitional affiliation. Thus, if this was true, race encoding was expected to decrease. Experiments (3) and (4) replicated experiments (1) and (2), except that the gender of the speakers varied, while race did not vary. If the hypothesis was true, the expected outcome was that gender encoding should not be affected by coalition encoding.

Results of experiment (1) showed that subjects were encoding a new dimension – coalition. The effect of race, however, was twice as large as the effect of coalition. In contrast, in experiment (2), when coalitions were marked by visual cues, the effect of coalition encoding increased (by more than 200%), while the effect of race decreased (by roughly 25%). Coalition membership encoding was higher than race coalition. Finally, in experiments (3) and (4), the encoding of gender was not significantly affected when coalitions were marked by visual cues.

Kurzban et al. (2001, p.15391) concluded from these studies that “the sensitivity of race to coalitional manipulation lends credence to the hypothesis that, to the human mind, race is simply one historically contingent subtype of coalition”.

4.3. Merits and problems

Results from Kurzban et al. [Kurzban et al. (2001), Cosmides et al. (2003)] seem to be good news: racialist classifications, on which racist cognition and behaviors depend, could disappear given some social re-engineering. We believe that their results suggest that concepts of race can be used by an evolved cognitive system whose function is to track coalitions. Thus, it may be true that the extent to which people classify according to race is affected by the extent to which race is believed to be a proxy for coalition. Nonetheless, we remain unconvinced by the main claim by Kurzban et al. that categorization by race is a by-product of this evolved cognitive system.
Their theory shares several problems with Hirschfeld’s work. Kurzban et al.’s ambitious theory relies only on two experiments with American undergraduates. No cross-cultural evidence is provided. Finally, they do not consider integrating their proposal with what we know about the cross-cultural diversity of the concepts of race.

Another shortcoming is that Kurzban and colleagues assume that races are coalitions [see also Hirschfeld (2001)]. It is not clear, however, that races – unlike nations or, sometimes, neighborhoods – really are cooperative groups [Gil-White (pers. Comm.)]. Black slaves in the United States came from various different cultures in Africa. White Europeans who went to Japan and China did not form a coalition either. Moreover, few people have put or are putting themselves at risk for the benefit of the members of their race, suggesting that races are not thought of as coalitions. Racist white Frenchmen may be willing to help other white Frenchmen, but not whites in general. Kurzban et al. could, however, reply that this objection misses the point. The claim is that because members of a race within a given society – say, Afro-Americans in the United States – form a cooperative group, people pay attention to racial membership. An individual uses race to categorize the individuals she meets, because around her, race is a cue to coalitional affiliation. The argument does not require blacks in general to form a cooperative group.

Notice however that this reply takes for granted that in some societies, races are coalitions [Kurzban et al. (2001)]. As we saw in the first section, however, there is evidence that racialism has repeatedly appeared during human history when humans with different phenotypes have met. For the sake of consistency, Kurzban and colleagues would have to say that in all these cases, races were coalitions. However, this would suggest that skin-deep features (like skin color, body shape, and hair appearance) bootstrap coalitional groups. And this fact would remain entirely unexplained by their theory. Hirschfeld’s account faces a similar problem. Ironically, like the social constructionists, Kurzban et al. as well as Hirschfeld fail to explain the prevalence of racialism across cultures and times. This cross-cultural and cross-historical recurrence of racialism needs to be explained.

Another difficulty is raised by the evidence that children think spontaneously about races in a biological way [Hirschfeld (1996)]. Kurzban et al.’s hypothesis does not explain this phenomenon. For it is not the case that all coalitions are thought of as animal species are: business firms and coalitions of nations, etc., are not thought of biologically.

Kurzban et al. could reply that races are thought of as coalitions and are thought of biologically. We pay attention to races, because they are coalitional groups, and we conceptualize them biologically. To evaluate this move, remember that what is at stake is the origins of the concept of race: which system produces this concept? Kurzban and colleagues’ experiments tentatively support the claim that concepts of races are used by a cognitive system that tracks coalitions. In societies where races happen to be very strong

\[12\] In reply, Kurzban (pers. Comm.) emphasizes that his work is aimed at falsifying social psychologists’ tenet that race is automatically encoded. For this purpose, a single experiment is sufficient. This is certainly an important aspect of Kurzban et al. (2001) and Cosmides et al. (2003). However, both papers put forward a positive claim, namely that categorization by race is a by-product of the coalitional cognitive system.
coalitions – which is arguably the case in many segments of the American society – racial membership may thus be used as a proxy for coalitional affiliation.

To support their claim that racial cognition is a by-product of the coalitional system, however, what matters is not how people use their concepts of races. What matters is how race membership is conceptualized. After all, it is plausible that this is determined by the cognitive system that produces the concept of race. However, Kurzban and colleagues have nothing to say about the biological content of humans’ concepts of race. Thus, Kurzban et al. fail to address the relevant question for explaining the origins of racialism. As a result, their findings do not support their claim that racial cognition is a by-product of an evolved coalitional system.

Finally, Hirschfeld’s, as well as Kurzban et al.’s, ultimate, evolutionary explanation of racial cognition is inadequate. Neither proposal considers the possibility that racialism could be a by-product of a cognitive system that would not be dedicated to small-scale social groups, but to different types of social groups. Hirschfeld (2001) even excludes this possibility. According to him, our ancestors lived in small, shifting coalitions. This is, however, an oversimplified view of the social life of our ancestors. It is important to distinguish different sorts of groups. These different types of groups may have created different evolutionary pressures and may have selected for different cognitive mechanisms.

Race membership seems indeed different from coalition membership. During human evolution, small groups were probably labile. As a result, categorization within these groups would plausibly require a constant updating. Being a member of this kind of group would not be conceived as being an inherent property. On the contrary, race membership is not conceived as being labile, but as being inherent.

We conclude that Kurzban and colleagues’ and Hirschfeld’s proposals are problematic. They assume that racial features bootstrap coalitional behaviors without explaining this. Kurzban and colleagues’ empirical evidence plausibly shows that races are sometimes taken as proxy for coalitions. However, this evidence falls short of showing that racial cognition is a by-product of the hypothesized coalitional cognitive system. To explain where racialism comes from, the content of our concepts of race has to be the main focus of inquiry. Finally, Hirschfeld’s and Kurzban et al.’s evolutionary hypotheses are unsatisfactory, for they fail to distinguish different kinds of groups that have raised specific evolutionary challenges during the evolution of human social cognition.

5. Is racialism a by-product of an evolved ethnic cognitive system?

5.1. “Ethnies” are not mere coalitions

Many anthropologists, particularly Boyd, Richerson, and their colleagues, have suggested that kin-based groups and coalitions are not the only evolutionary important social groups. Our ancestors have also belonged, for an evolutionary significant time, to larger groups, called “tribes” or “ethnies” [Richerson and Boyd (1999), Boyd and Richerson (2001, 2004)].
The notion of ethnie used here picks out large groups that consist of several thousands of individuals – for instance, the Nuer in Sudan. Ethnies are divided into smaller units, sometimes called “bands”. They form cultural units. Many culturally transmitted norms, including norms of cooperation, are recognized by all the members of a given ethnie and these norms differ from the norms that prevail in other ethnies [Richerson and Boyd (1998, 1999)]\textsuperscript{13}. Few other properties are common to all ethnies [Knauff (1991)]. This type of social organization is specifically human. The first ethnies appear in the archaeological record 50,000 years ago [Klein (1999)] and may have existed earlier [but see Knauff (1991)]\textsuperscript{14}.

Ethnies are also characterized by ethnic markers [Richerson and Boyd (1999), McElreath, Boyd and Richerson (2003)]. Coethnics use various symbolic signs (e.g., body paintings, clothes, jewels, accents etc.) to signal their ethnic membership. These markers vary across ethnies. Signs that may have functioned as ethnic markers are well attested in Europe 40,000 years ago and may have existed much earlier in Africa [McBrearty and Brooks (2000)].

It has been hypothesized that this form of social organization has created \textit{sui generis} selective forces [Richerson and Boyd (1998, 1999), Boyd and Richerson (2001)]. According to Boyd and Richerson, our ancestors have evolved to be emotionally committed to the norms of their ethnies. Ethnies may thus have been an environment that produced new selective pressures. These pressures may have selected for an ethnic cognition beside our kin and coalitional cognition. Gil-White suggests that this is the key for understanding racial cognition.

5.2. An adaptive scenario: Ethnic cognition and the exaptation of human folk biology

According to Gil-White, humans have evolved an ethnic cognitive system that is based on our folk biology [Gil-White (1999, 2001a)]. That is, humans have evolved to think of ethnies as if they were biological species. Our evolved folk biology contains the innate knowledge about biological species, including a belief in essences, and the reasoning biases that are applied to species [Atran (1990), Medin and Atran (1999, 2004)].

Ethnies triggered our ancestors’ folk biology, because they have several important characteristics in common with species [Gil-White (2001a)]. First, ethnies are characterized by clusters of stable, culturally transmitted behavioral norms, and norms tend to vary across ethnies [Henrich and Boyd (1998)]. Thus, like conspecifics, coethnics behave similarly, and members of different ethnies behave differently, like members of different species. Besides, interactions across ethnic boundaries have a lower payoff than interactions within these boundaries, since social norms vary across ethnies. Our ancestors were probably sensitive to these costs [McElreath et al. (2003)]. Thus, norm

\textsuperscript{13} Of course, some norms are specific to subethnic groups and others are common to several ethnies. Ethnies should not be thought of as homogenous, isolated groups that differ linguistically and culturally \textit{in toto} from other ethnies [Richerson and Boyd (1999)].

\textsuperscript{14} We will not review the evidence for this hypothesis [Bettinger (1991), Rodseth et al. (1991), Richerson and Boyd (1998, 1999), Boyd and Richerson (2001), Richerson, Boyd and Henrich (2003)].
boundaries tended to coincide with social interactions, particularly with mating preferences. As a result, endogamy and descent-based ethnic membership were probably prevalent. Finally, our ancestors tended to broadcast their ethnic membership. Parents and children usually displayed the same markers, which was similar to a species-specific morphology.

This application of our folk biology to ethnies is an exaptation (and not a misfiring). A higher reproductive success was conferred upon individuals whose folk biology was easily triggered by ethnic markers. For thinking about ethnies as if they were species may be good epistemology—though it is certainly bad science. This belief fosters inductive generalizations on the basis of limited contacts. Since members of other ethnies have many behaviors in common, such wild generalizations may tend to be true. If our ancestors occasionally interacted with members of other ethnies, this inexpensive learning strategy may have been adaptive. Moreover, a biological view of the ethnic world plausibly reduces the frequency of interactions across ethnic boundaries whose success requires shared norms; which is especially true for mating [Gil-White (2001a)]

According to Gil-White, races trigger by mistake our ethnic module. For the physical properties that define race membership are similar to some ethnic markers. Moreover, the physical properties are shared by parents and children, like ethnic markers are.

5.3. Empirical evidence

Gil-White studied Mongols and Kazakhs. Both ethnies are semi-nomadic pastoralists. They have the same social standing and live in the same environment, although they are territorially segregated. In both groups, children inherit their ethnicity from their father [Gil-White (2001a)].

Gil-White (2001a, 2002) relied on the “switched-at-birth” experiment that was developed by Gelman and Wellman (1991) and used by Hirschfeld (1996) and Solomon et al. (1996). Briefly put, subjects were told that a child, whose parents were Kazakh, was raised by Mongols (or vice-versa). They were asked what the ethnicity of the child is. Gil-White aimed to determine whether ethnic ascription was sensitive to the rearing environment or was transmitted by descent.

Roughly two-thirds of the subjects concluded that the child was Kazakh. Thus, a large majority of Gil-White’s subjects expected ethnic membership to be impervious to the rearing environment. According to Gil-White, these cross-cultural results support the claim that human ethnic cognition is essentialist.

15 Of course, migrations, cultural influences, and economic exchanges occur between ethnies [e.g., McBrearty and Brooks (2000)]. Interethnic alliances are also known. Exchanges across ethnic boundaries, however, are very different from exchanges between coethnics.

16 Gil-White’s theory also predicts that endogamous and descent-based groups will trigger this hypothesized ethnic cognitive system [Gil-White (2001a)].

17 Scupin (pers. Comm.) has run a similar experiment with university students from Thailand. The results are mixed. Scupin used two sets of questions. For the first set of questions, only 50% of the students gave essentialist answers in the “switched-at-birth” experiment.
Gil-White (2001a) investigated further. He tried to see whether people associate ethnic membership with the possession of an essence. As we saw, essentialist thinking posits that category members share a hidden essence that determines what they really are. Essentialists should thus deny that the child could ever become exactly like her adoptive parents. Most of the subjects who gave essentialist answers asserted that the child would be somewhat physically similar to and behave like the members of the ethnie of her biological parents. More than half of them replied that the blood of the ethnie of her biological parents could never be “erased.” This is seen as confirming the essentialist prediction. According to Gil-White, these replies show that descent-based ethnic membership is more than a mere labeling of one’s ascent. It implies developing some bodily and behavioral properties of one’s ethnic group and receiving a limited influence from the rearing environment.

Moreover, he discovered in informal discussion that some subjects who gave nonessentialist answers held some beliefs that were at odds with their explicit nonessentialist assertions. For example, some claimed, on the one hand, that the child would belong to his rearing ethnie and, on the other hand, that she would be unable to do some things that were characteristic of that rearing ethnie (e.g., sorcery). This second answer was taken to imply a commitment to essentialism.

Finally, Gil-White (2001a) replaced ethnic names with the names of clans, i.e., subethnic groups, in the questionnaire presented above. Crucially, the answers differed markedly from the answers triggered by the ethnic questionnaire.

5.4. Merits and problems

Gil-White’s theory is not entirely new. He endorses the most convincing points of Hirschfeld’s psychological account and Boyd and Richerson’s picture of the evolution of human social behavior and cognition. To our knowledge, he is the first, however, to synthesize these disparate ideas, to provide a rather compelling evolutionary hypothesis, and to build a seductive explanation of racialism based on this. We are convinced by many aspects of Gil-White’s theory, particularly by the hypothesis of an evolved ethnic cognitive module. We also believe that he made a convincing case that this system underlies racialism. Nonetheless, several problems remain.

First, Gil-White’s theory shares several problems with Hirschfeld’s theory. Gil-White interprets the empirical evidence through the psychological essentialism framework, although a belief in essences is not necessary to explain the reasoning patterns he presents. Moreover, his proposal does not cast much light on the interaction between culture and our evolved ethnic system. We also disagree with the way Gil-White deals with the diversity of people’s answers to his questionnaire. He purports to explain away these differences, claiming that deep down, everyone is committed to a biological view of ethnic membership, and that nonessentialist models are superimposed on this biological view. Since differences are explained away, no attempt is made to elucidate why some individuals would not conceptualize ethnic membership biologically. Gil-White certainly provides some circumstantial evidence that some nonessentialist individuals believe that ethnic membership is an inherent, descent-based
property. He does not provide, however, any systematic evidence that all or even most nonessentialists share this belief. Given that many people deny having a biological view of ethnic membership, we feel that the evidence falls short of establishing Gil-White’s bold claim.

Second, it has to be recognized that the evidence for Gil-White’s proposal is scanty. It is supported by Hirschfeld’s developmental data and by his own anthropological data. Gil-White (1999) has tested his hypothesis against the ethnographic record. Particularly, he proposes that the evidence does not support the claim that in some cultures, people can change their ethnic and racial identity at will, which would falsify his views. This survey of the ethnographic record, however, does not establish that people think about ethnic membership biologically. To support this idea, it must be shown cross-culturally that we reason similarly about species on the one hand, and about races and ethnicities on the other hand. Thus, it seems fair to conclude that regarding Gil-White’s proposal, the jury is still out. More cross-cultural evidence is needed.

Third, Gil-White proposes that our evolved folk biology is a skeletal commitment to essentialism that is filled in by culturally transmitted biological beliefs [Gil-White (2001a, pers. Comm.)]. That is, children believe that species are characterized by an essence, but have no specific belief about the nature of this essence or about its transmission. Beliefs about these matters are learned. We agree with Gil-White that most folk biological beliefs are culturally transmitted.

There remain, however, two points of contention between us and Gil-White. First, there is some convincing evidence that we are disposed to think of species in a hierarchical way [Atran (1990), Malt (1995), Medin and Atran (1999, 2004)] and to reason about them in a peculiar way [for a general review, see Medin and Atran (1999, 2004)]. We may also have evolved to expect species-specific properties. But our evolved folk biology may consist of little more. Particularly, essentialism – that is, a belief in essences – may be culturally transmitted and culturally specific [but see Gelman (2003), Medin and Atran (2004)]. Like Hirschfeld (2001), Gil-White may thus assume a folk biology whose evolved content is too rich.

Moreover, Gil-White suggests that we are predisposed to learn specific folk-biological beliefs, for example, the beliefs that conspecifics naturally mate with each other and that they engender conspecifics [Gil-White (2001a, pers. Comm.)]. We are not convinced. The adaptive function of these beliefs is far from obvious. Certainly, once animal domestication was included in humans’ behavioral repertoire, these beliefs were useful. However, animal domestication is a very late phenomenon in human evolution. Thus, it is not clear that we have evolved to entertain them. Gil-White could reply that the universality of these beliefs supports his hypothesis. This is not necessarily so, however. Everybody believes that the sky is blue, because it is a true belief that is easily learned by each individual. Similarly, universal folk biological beliefs may simply be true beliefs that are easily learned either by individuals or by cultures. There is no need to suppose that we have been disposed by design to learn them. Moreover, there is some cultural variation concerning some of these beliefs. In some cultures, people believe in the possibility of cross-species fertile reproduction [Atran pers. Comm.].
Fourth, if ethnies and races are thought of as being like animal species, then they are thought of as being peculiar species. Interethinic reproduction (e.g., rapes during wars) has probably existed during human evolution and was common for known ethnies. Interracial marriages have been common during human history. Gil-White’s proposal does not explain this fact [Astuti (2001), Boyer (2001b)]. In reply, Gil-White (2001a) suggests that when species are similar, e.g., donkeys and horses, we do not have the intuition that mating is impossible, simply that it is not natural. It is far from clear, however, that during war between ethnies, raping women or kidnapping them in order to mate with them is spontaneously viewed as unnatural. Our folk theories of species, ethnies, and races may also differ in other respects. For example, we conceptualize species hierarchically [Atran (1990), Medin and Atran (2004)]. Ethnies, however, do not seem to be thought of as belonging to hierarchies. More should be said about the differences between these three folk theories.

We conclude that Gil-White has shown that we plausibly evolved an ethnic cognitive system and that it underlies racialism. His views assume, however, an unlikely rich evolved folk biology. Finally, the integration of his approach with the social constructionist evidence remains to be done.

6. Conclusion

Most contemporary theories of racialism are inspired either by social constructionism or by a cognitive-cum-evolutionary approach. There has been little contact between proponents of these two approaches. To overcome the prevalent theoretical tribalism and inspire integrative theories of racialism, we have reviewed some significant contributions to the recent literature on racialism [see also Machery and Faucher (in press)].

We have first presented the theoretical tenets and some results of the social constructionist research tradition on racialism. The way people conceptualize race membership is strongly influenced by their cultural niche. We have, however, underscored that the similarities between culture-specific concepts of race remain unexplained by this approach. These similarities suggest that racialism results from a universal cognitive system. Three cognitive-cum-evolutionary theories were then reviewed: all see racialism as a by-product of an evolved cognitive system. Despite several shortcomings, Hirschfeld’s contribution to the understanding of racialism is important. Although more evidence is needed, he has made a convincing case that racialism results from a domain-specific cognitive system, that it develops early, and that races are thought of in many respects as

---

18 This fact is also problematic for Atran’s view [see also Rothbart and Taylor (1992), Atran (2001)]. According to Atran, there is no need to posit any new cognitive system to explain racial cognition: races merely trigger our folk biology because of their perceptual properties. He would have to explain the differences between human racialist and biological cognitions. Positing an ethnic cognitive system may be instrumental in explaining these distinctions, even if this system is based on an exaptation of our folk biology.
animal species are. However, we have rejected his human kind module and his evolutionary hypothesis. Kurzban et al. have shown that an evolved coalitional cognitive system may use concepts of races. However, we have rejected the claim that racialism results from this cognitive system. Finally, we have presented Gil-White’s theory: humans have evolved an ethnic cognitive system and this system underlies racialism. Despite several shortcomings, we have underscored the merits of this proposal.

On the basis of this critical review, we can put forward 11 requisites for future theories of racialism. Some are uncontroversial, while others reflect our own theoretical commitments.

1. Races do not exist. Thus, there is no evolved module for racialism.
2. Racialism varies across cultures. A theory of racialism has to accommodate this diversity.
3. Classifications on the basis of phenotypic properties are similar all over the world. These similarities ought to be studied in more detail by historians and anthropologists. A theory of racialism has to account for them.
4. These similarities suggest that racialism is the product of a universal cognitive system. Researchers have to specify precisely the cognitive mechanism(s) that underlies racialism.
5. To identify the origins of the concept of race, the cognitive system that produces the concept of race should not be confused with the cognitive systems that use racial identification.
6. A theory of racialism should account for the evolution of this cognitive system. It should be consistent with the best paleoanthropological theories of the evolution of human social behavior.
7. Researchers should not interpret the empirical evidence through the controversial framework of psychological essentialism.
8. Researchers should not assume that our evolved folk theories, e.g., our evolved folk biology, are rich and rigid. Most of our folk beliefs are culturally transmitted.
9. Theories of racialism tend to be ambitious. In contrast, empirical evidence tends to be scanty. Psychologists should derive many predictions from their theories. Various experimental paradigms should be used to test these predictions.
10. Cross-cultural data are needed to establish the universality of patterns of reasoning etc.
11. Psychologists should pay attention to individual differences. They may cast some light on the mechanisms that underlie racialism.

Although no current theory of racialism satisfies all these requirements, our review suggests that recent theories have significantly increased our understanding of racialism. We would like to thank François Athané, Katarzyna Buczek, Francisco Gil-White, Lawrence Hirschfeld, Robert Kurzban, Ray Scupin, and Dan Sperber for answers to our questions, access to previously unpublished material, and comments on previous drafts of this article.
References

Scupin, R., “Muslim identity and nationalism in Buddhist Thailand”, unpublished manuscript.
Chapter 47

NEUROSEMANTICS AND CATEGORIES

CHRIS ELIASMITH

Department of philosophy,
University of Waterloo, Waterloo

Contents

Abstract 1036
1. Introduction 1037
   1.1. Why “neuro”? 1037
   1.2. The explanandum 1039
2. Mental representations as neural codes 1040
   2.1. Representations 1040
   2.2. Transformation 1041
   2.3. A representational hierarchy 1042
3. The meaning of neural representations: Neurosemantics 1043
   3.1. The representation relation 1043
   3.2. A neurosemantic theory 1044
      3.2.1. Systems 1044
      3.2.2. Vehicles 1045
      3.2.3. Referents 1046
      3.2.4. Content 1047
   3.3. Discussion 1049
4. Misrepresentation 1050
5. Conclusion 1052
References 1052

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved

1035
Abstract

A theory of category meaning that starts with the shared resources of all animals (i.e., neurons) can, if suitably constructed, provide solutions to traditional problems in semantics. I argue that traditional approaches that rely too heavily on linguistics or folk psychological categories are likely to lead us astray. In light of these methodological considerations, I turn to the more theoretical question of how to construct a semantic theory for categories informed by neuroscience. The second part of the chapter is concerned with describing such a theory and discussing some of its consequences. I present a theory of neural representations that describes them as a kind of code, and show that such an understanding scales naturally to include complex representations such as concepts. I use this understanding of representational states to underwrite a theory of semantics. However, the theory must be supplemented by what I call the statistical dependence hypothesis. Content is then determined by a combination of the states picked out by this hypothesis and the neural decoders that define subsequent transformations of the neural representations. I briefly describe a solution to the traditional problem of misrepresentation that is consistent with this theory.
1. Introduction

As described in many contributions to this volume, categories are concepts, and concepts are representations. Some claim that these representations are learned by individuals, others that they are socially constructed, and still others that they are innate. There are theories of perceptual categorization, categorization in data mining, and the metaphysics of categorization. There are contributions relating to categorization in computer science, philosophy, psychology, and neuroscience. In most such discussions, there are fundamental assumptions that usually remain implicit regarding categories. These have to do with the meaning, or semantics, of categories.

Of course, the semantics of categories has been discussed at length here and elsewhere [Loar (1981), Lycan (1984), Millikan (1984), Dretske (1988), Fodor (1998)]. However, what is unique about the discussion in this chapter is the perspective that I adopt: I address semantics from the perspective of computational neuroscience. Most attempts at addressing semantics have been linguistic or psychological and rely on more traditional accounts of computation. I argue that adopting a neuroscientific view has a number of advantages. And, to demonstrate this concretely, I outline a theory of category meaning based on recent work in computational neuroscience. To show that this theory has some advantages over past theories, I apply it to the traditional problem of misrepresentation. That is, the problem of determining how we can miscategorize entities in the world. Somewhat surprisingly, the problem of misrepresentation is simply ignored by many empirical approaches to categorization. But it is, in fact, the very central problem of determining what the appropriate labels for categories actually are.

1.1. Why “neuro”?

Many philosophers who have proposed semantic theories have focused on the propositional content of beliefs and language [see, e.g., Loar (1981), Evans (1982), Harman (1982), Lycan (1984), Block (1986), Fodor (1998)]. This project has been less than obviously successful. As Lycan (1984), a proponent of the approach, has put the point:

Linguistics is so hard. Even after thirty years of exhausting work by scores of brilliant theorists, virtually no actual syntactic or semantic result has been established by the professional community as known. (p. 259)

But Lycan, like most researchers in the field, is determined to continue with the project using the same methods, and shunning others: “And there must be some description of this processing that yields the right predictions without descending all the way to the neuron-by-neuron level” [Lycan (1984) p. 259]. After significantly more than 30 years of difficulty, it seems rather likely that those neuron-by-neuron details actually do matter to a good characterization of the syntax and semantics of mental representations (and, through them, language).

There are reasons other than a simple lack of success for language-based approaches to think that neuroscience may be a better starting point for such theories. For one thing,
participants in the debate generally agree that mental content should be naturalized. That is, meaning deserves a scientific explanation that refers to objects found in nature. Linguistic objects, like words, are presumably one kind of object found in nature. But, it is a mistake to give an explanation of meaning in terms of words and complex natural language, since this is to explain one poorly understood natural concept (mental meaning) in terms of another (linguistic meaning). In fact, such an explanation would be perfectly circular if we were giving an explanation of meaning that relied on the content-carrying capacity of words. For this reason, most language-based explanations try to introduce other notions as primitive, notions like “information,” “cause,” or “function.” But even with these less circular primitives, assumptions regarding the structure of mental meanings (e.g., that they are language-like in various ways) are often deeply influenced by the example of natural language.

This is a serious problem because language is only a tiny domain of the application of the notion of mental meaning. Language, as most linguists understand it, is a human specialization. Thus it is unique to one species among billions. This is a good reason to think that starting with language, or even focusing on language, when constructing a theory of content is a dubious tactic. This is true unless we have prima facie evidence that most nonhuman animals do not have internal representations. We have no such evidence, and, in fact, there is significant evidence to the contrary [Redish (1999), Eliasmith (2003)]. At a minimum, the use of symbols for communication is rampant in the animal kingdom. Bee dances, monkey calls, whale songs, bird songs, etc. are all instances of animals communicating properties of the environment via symbols that refer to the things having those properties. So it is much more common for there to be meaning without language than meaning with language. Meaning, it seems, is prior to language.

Further support for this contention comes from the clear evidence that language is not necessary for mental meaning. People who have had the misfortune of growing up without natural language, but later learn language, are able to recall events that precede their linguistic competence [Garmon and Apsell (1997)]. So we need a theory that can account for meaning regardless of the presence of natural language.

In addition, if we take linguistic capacities to be the result of a somewhat continuous evolutionary process from less complex to more complex organisms, then the fact that language is a human specialization suggests that it is a far more complex phenomenon than “merely” having neural states with meaning. Even those, like Chomsky (1986), who think that language is a specifically human ability that does not have evolutionary precursors, argue that language is particularly complex. Being able to deal with the complexities of language suggests a uniquely powerful set of computational abilities in humans. To begin our explorations of meaning by examining a phenomenon found solely in the most complex exemplar systems with meaning is, practically speaking, a bad tactic [Bechtel and Richardson (1993)]. This might serve to explain the lack of progress noted by Lycan.

Given these considerations, it should be clear just how bad an idea it is to hang our theories of mental meaning on what we know about the structure of natural language. To claim
that animals have a “language of thought” simply because our theory of meaning depends on our understanding of human language is quite clearly a claim of the tail-wagging-the-dog variety [see, e.g., Fodor (1975)]. In contrast, a theory of meaning that starts with the shared resources of all animals (i.e., neurons) can, if suitably constructed, draw the anticipated boundaries in virtue of principled differences in the use or organization of these resources. So starting at the “neuron-by-neuron” level may not be such a bad idea after all.

1.2. The explanandum

In the last 20 years or so, there has been a plethora of philosophical theories trying to provide naturalistic explanations of mental meaning. They have included covariance theories [Dretske (1981, 1988), Fodor (1981, 1998)], functional role theories [Harman (1982, 1987), Block (1986)], adaptational role theories [Millikan (1984), Dretske (1988, 1995)], and isomorphism theories [Cummins (1996), O’Brien and Opie (2004)]. It is not always clear that these theorists have the same target in mind when providing their explanations. So, to begin, I outline what I take to be expected of a good, naturalistic theory of meaning.

For the most part, contemporary naturalistic theories of mental meaning are part of a tradition concerned primarily with linguistic meaning. In this tradition, the meaning of a sentence is the abstract proposition that the sentence expresses. For mental states, it is the thought that has this same meaning (or content). Simple sentences, like *The star is bright*, have traditionally been analyzed as an ordered pair of elements <object, predicate> whose meanings are thought to be either senses (sometimes called the “Fregean” view) or references (sometimes called the “Russellian” view).

In mental, or psycho-, semantics, however, it is not clear what the object/predicate distinction should amount to (unless one presupposes a Language of Thought, which I do not). This is because psychosemanticists have been much more concerned with the meaning of simple mental representations, i.e., categories or concepts, which, while they carry content, are often taken to be the *constituents* of sentences. Concept labels, like *dog*, do not explicitly declare objects and predicates. However, concepts are often taken to be category labels, whose members might share some common or “core” properties. And this category/property distinction is often employed like an object/predicate one (e.g., to say “My concept [DOG] includes mostly furry objects” is to say that “Dogs are furry”).

For present purposes, this parallel means that doing neurosemantics is really doing semantics in the linguistic/psychological tradition [which some, like Fodor (1987), may strongly doubt]. This is because I take the meaning of a neural representation to be what that representation tells you about what it represents. An equivalent way of saying this is that the meaning of a neural representation is the set of properties ascribed to something by that representation. So, I too adopt an analogous object/predicate or category/property

---

1 Although I take it that, in general, the “something” can be objects, events, relations, and so on, here I will concentrate only on objects. As well, I will concentrate on predicate-object pairs only (e.g., not compound sentences).
distinction. However, this distinction can be “subpsychological” (e.g., the firing of a retinal neuron ascribes the property of a certain luminance at a retinal location). As a result, I take my explanandum to be shared with these past theories. And I take that explanandum to be: How is the set of properties ascribed to an object by its representation determined?

Before addressing this question directly, in Section 3, the next section takes a brief detour into recent work in computational neuroscience that sheds light on the kinds of neural states that are relevant for mental representation [Eliasmith (2003)].

2. Mental representations as neural codes

2.1. Representations

As discussed in detail in Eliasmith and Anderson (2003), it is possible to adapt a standard information theoretical account of codes to understanding representations in neural systems. Codes, in information theory, are defined in terms of a complementary encoding and decoding procedure between two alphabets. Figure 1 depicts a toy case of a code, where $B$ would be said to be a code for, or representation of, $A$. In this case, both alphabets are binary digits.

In order to characterize representation in a neural system in a similar way, we need to identify each of these procedures and their relevant alphabets. The encoding procedure is straightforward: it is the transduction of stimuli by the system resulting in a series of neural action potentials. Encoding is what neuroscientists typically talk about. For example, when we present an image of a dog to a subject, some neurons or other “fire”\(^2\). Unfortunately, neuroscientists often stop here in their characterization of representation (i.e., they want to claim that those neurons represent a dog), but this is insufficient (stopping here is equivalent to adopting a naive causal theory). We also need to

\[
\begin{align*}
A & \quad f(A) = B \\
10110 & \quad 01001 & \quad 10110
\end{align*}
\]

Encoding Decoding

\[
f(x \in A) = \begin{cases} 
0 & \text{for } x = 1 \\
1 & \text{for } x = 0
\end{cases} \quad f^{-1}(x \in B) = \begin{cases} 
0 & \text{for } x = 1 \\
1 & \text{for } x = 0
\end{cases}
\]

Fig. 1. A simple example of an encoding/decoding relation.

\(^2\) The precise nature of this encoding has been well quantified [Bower and Beeman (1998)].
identify a decoding procedure; otherwise, there is no way to determine the relevance of
the encoding for the subject. If no information about the stimulus can be extracted from
the spiking neurons, then it makes no sense to say that it represents the stimulus. After
all, representations should be able to “stand in” for that which they represent. As dis-
cussed in Eliasmith and Anderson (2003), there are methods for determining good, lin-
ear decoders that decode time-varying, distributed population representations.

Having specified the encoding and decoding procedures, we still need to specify the
relevant alphabets. Neuroscientists generally agree that the basic element of the encod-
ing alphabet is the neural spike. However, there are many possibilities for how such
spikes carry information. Of these possibilities, arguably the best evidence exists for a
combination of timing codes and population codes [see Abbott (1994) for an overview;
Salinas and Abbott (1994), Rieke et al. (1997)]. Thus, I take the encoded alphabet to be
the set of temporally patterned neural spikes over populations of neurons.

Unfortunately, it is more difficult to be specific about the nature of the decoded alpha-
bet. Nevertheless, a justifiable assumption is that what is encoded (at least initially) are
physical properties in some specifiable units (e.g., m/s, kg, etc.). More complex proper-
ties (e.g., red, hot, conspecific, etc.) are inferred on the basis of representations of prop-
erties with identifiable units. As a result, the units of such properties will also be
complex, and are currently undetermined (though empirically determinable).

2.2. Transformation

A representational characterization is of little use if it does not play a role in under-
standing the function of a neural system. Conveniently, the preceding characterization of
neural representation can be easily extended to account for neural transformations. This
is because, like representations, transformations can be characterized using the notion of
decoding. Specifically, related methods exist to find a “transformational decoder” which
extracts information other than what the population is taken to represent [Eliasmith and
Anderson (2003)].

Given this understanding of neural computation (i.e., as encoding and (transforma-
tional) decoding), a problem arises because the information encoded in a population
may now be decoded in a variety of ways. Since representation is also defined in terms
of encoding and decoding, it seems that we need a way to determine which of the set
of possible decodings is the relevant one for defining the “true” representation of the
original population.

To resolve this problem, I have elsewhere specified that what a population represents
is determined by the decoding that results in the variable that all other decodings are
functions of [Eliasmith (2003)]. This does not completely resolve the difficulty since
any given variable can be written as a function of another variable that is a function of
it. This remaining difficulty can be resolved by noticing that the relevant variable is that
variable whose units are part of a coherent and useful theory [Eliasmith (2003)]. This
merely amounts to the observation that our neuroscientific theories ought to be consis-
tent and continuous with our other physical theories. For instance, supposing that $x$ is
an object’s position in the world, and that \( x \) and \( x^2 \) are decoded from some encoding in a neural population, we would claim that the population represents position because all uses of the representation can be related to position and “position” is a standard category in the sciences (whereas “position squared” is not). In other words, we should (at least initially) take cognitive systems to represent the properties that we describe the world in terms of since cognitive systems represent the world. This is not a demand for realism regarding all properties in our current descriptions of the world, but a demand for consistency between those descriptions and descriptions of cognitive systems.

2.3. A representational hierarchy

For illustrative purposes, let us consider a neural population’s representation of the horizontal position of an object; such a representation is found in the lateral intraparietal cortex [Andersen, Essick and Siegel (1985)]. This representation can be understood as the encoding of a scalar variable into a series of neural spikes\(^3\). The units of that variable are “degrees from midline.” Using the quantitative tools mentioned earlier, it is possible to determine the representational decoder. Once we have the decoder, we can then estimate what the actual position of the object is, given the neural spiking in this population. Comparing this result to an independent characterization of the object’s position, we can then determine precisely how well the original property (or specific aspects of it) is represented by the neurons in the population.

This simple example not only describes how to characterize representation, it also shows how we can move from talking about single neuron responses to talking about “higher-level” variables, such as “horizontal object position.” That is, we can move from discussing the “basic” representations (i.e., neural spikes) to “higher-level” representations (i.e., mathematical objects with units). As I have discussed elsewhere, even such a simple example as this demonstrates how to build up a “representational hierarchy” that permits us to move away from a “neuron-by-neuron” description, while remaining responsible to it [Eliasmith (2003)]. To continue up the hierarchy, we could talk about a larger population of neurons that encodes object position in three-dimensional space. We could then spell out the details of lower levels of the hierarchy by dissecting this higher-level description into horizontal, vertical, and depth positions, and then dissect these descriptions further into neural responses. Which description we employ will depend on the kind of explanation we need and the specific properties of the neural system we are describing (e.g., the nature of the neurons’ tuning curves).

Notably, this hierarchy can be rigorously defined to include scalars, vectors, functions, and any combinations of these [Eliasmith and Anderson (2003), p. 63]. Because all of the levels of this hierarchy can be written in a standard form, it is natural to suggest that it provides a unified way of understanding representation in neurobiological systems. The wide range of mathematical objects that can be related to neural processing in this way supports the stronger claim that this hierarchy is general enough to

\(^3\) Although this understanding is oversimplified and unrealistic, it is a useful approximation for this purpose.
capture all neural representation. That is, since all mental representations can be described as some combination of scalars, vectors, and functions, and those mathematical objects can be neurally represented using these methods, these methods can be used to describe all mental representations.

That being said, it is important to keep in mind the fact that there can be more than one possible implementation at each level of the hierarchy (e.g., more than one way a neural system can represent a three-dimensional vector). So this hierarchy in no way relieves us from being responsible to the empirical facts regarding a particular neurological system. Nevertheless, this representational hierarchy can unify our understanding of representation in neurobiological systems from neural firings to psychological-level representations.

Thus, regardless of what kind of mathematical objects with units higher-level representations turn out to be, the previous method for understanding neural representation applies. While adopting the representational hierarchy has definite consequences for the form of a preferred representational account, it is silent as to which particular one is correct (i.e., it does not determine which mathematical objects are appropriate for which aspects of neural function). This is a desirable result, since we simply do not know what the right higher-level representations are at this point in the history of neuroscience. While there is general agreement on the mechanism underlying individual neuron activity, there is little agreement on how populations of neurons manage to represent the world. Presumably, the right account will be the most coherent and predictively successful one – an adjudication that it is simply too early to make.

3. The meaning of neural representations: Neurosemantics

3.1. The representation relation

To this point, I have only addressed how to characterize the kinds of states that can carry content. In particular, I have done so using the tools of computational neuroscience. I have not, however, addressed how to determine the content carried by those states.

Standard accounts of the representation relation, including both causal and conceptual role theories, identify a three-place relation in describing representation: for causal theories, there is the representation, the thing it represents, and the context under which it is a representation (and not just an effect). For conceptual role theories, there is the representation, the thing it represents, and the role it plays (i.e., its context as defined by the system of concepts).

---

4 This may seem implausible if we are concerned that canonical cognitive representations such as images and sentences are not mathematical objects of these kinds. However, note that any digitizable representation can be considered as a vector, and functions naturally account for analog representations.

5 For Fodor, such a context is defined by the nomic relations that obtain, in particular, those that have asymmetric dependencies and those that do not. For Dretske, the context is determined through an evolutionary and learning history.
One of the reasons that this three-place relation is so ubiquitous stems, I suspect, from the nearly universal commitment among philosophers and neuroscientists to understanding neurobiological systems as information processing systems [see, e.g., Dretske (1981), Bialek and Rieke (1992), Van Essen and Anderson (1995), Rieke et al. (1997), Fodor (1998), Koch (1998), Eliasmith and Anderson (2003)]. Formally, the information relation is a three-place relation: A \{channel\} carries \{information\} with respect to a \{receiver\} [Reza (1994) p. 2]. These three places are necessary and sufficient for an adequate and general definition of Shannon and Weaver-style information [Shannon (1948/1949)]. Also notice that they loosely align with the three places of the representation relation as described above (i.e., channels or vehicles carry information or content with respect to receivers or systems). It is not surprising that both the representation relation and the information relation are three-place relations since the nature of the latter often informs intuitions about the nature of the former, given such a commitment.

However, the metaphor relating mental representation and information does not hold up under scrutiny. In particular, as Dretske (1983) has been at pains to point out, there is no such thing as “misinformation” in the same sense as there is “misrepresentation.” That is, information (in the technical sense) is never wrong about anything. Representations, by contrast, can be. This disanalogy is important because it highlights the need to identify a fourth element in the representation relation: the referent, that is, the object or event that the content of the representation is supposed to be about. The importance of this element, of course, is something like what Frege wanted to highlight with his distinction between reference and sense (although it is not the same).

I take it, in fact, that failure to distinguish contents and referents is what leads to many of the difficulties for both causal and conceptual role theories. Both equate contents and referents, though by slightly different routes. Specifically, causal theories take referents to be contents, with the result that surreptitiously changing referents leads to sometimes counterintuitive shifts in meaning and difficulties explaining misrepresentation. Conceptual role theories, in contrast, are in a position where it seems they must take contents to be referents, not allowing them to explain the relevance of truth conditions to meaning.

Consequently, I adopt the following four-place schema for the representation relation (the “representation schema”):

A \{vehicle\} represents a \{content\} regarding a \{referent\} with respect to a \{system\}.

3.2. A neurosemantic theory

Given this representation schema, what a semantic theory must do is explain what the four constituting theoretical objects are. The following description is based on work presented in Eliasmith (2000).

3.2.1. Systems

Although I call the fourth relatum the “system”, as I take this formulation of the representation relation to be general, philosophers of mind tend to narrow the scope of this
schema to include only natural biological systems (often just human beings). In other words, they are not interested in representation writ large, but only in neurobiological representation. I adopt this narrower perspective as well. So, the fourth element of the schema will be, for my purposes, “the (complete) nervous system.” This is in contrast to understanding “system” as either a more general physical system or an abstract representational system [see, e.g., Goodman (1968)].

3.2.2. Vehicles

Vehicles are the internal physical objects, or “representations,” that carry representational contents. As evident from Section 2.3, I take there to be two kinds of vehicles: basic and higher level. Basic vehicles are neurons, understood as functional units. Higher-level vehicles are simply sets of neurons.

To get a better understanding of what this distinction entails, consider a debate in neuroscience regarding whether neurons in visual area V1 are “edge detectors” or “orientation filters” [Van Essen and Gallant (1994)]. This example raises a concern about the distinction I have drawn between basic and higher-level vehicles. Notice that the terms “edge detectors” and “orientation filters”: (1) are descriptions of single neurons; and (2) pick out vehicles based on their contents.

This first point raises the concern that taking basic vehicles as uncontentiously identifiable with single neurons is simply mistaken. Since there are debates over such terminology, it seems clear that the identities of basic vehicles really are not uncontentious after all. However, I take it that neurons are basic vehicles only as functional units, not as carriers of content. That is, it has been well established that single neurons have certain physical properties that result in predictable system-independent functioning. The plethora of highly detailed single-cell models attests to this fact [e.g., Bower and Beeman (1998)]. However, it is not well established what the behavior of such single neurons means. This leads us to the second point.

The reason that the identity of vehicles is not easy to settle is that vehicles are often named after the contents they supposedly carry. “Edge detectors” are so called because they are thought to be activated by edges and used to detect edges. “Orientation filters” are so called because they are thought to be activated by spatial orientations and used to analyze orientation gradients in visual images. In some ways, this seems unnecessarily confusing, but it demonstrates how important content determination is to theories in neuroscience as well as philosophy. In any case, once we keep in mind that the naming practices of vehicles have a tendency to rely on content claims, claims about the nature of vehicles (simpliciter) become less contentious. So, saying that basic vehicles are neurons as functional units and higher-level vehicles are sets of neurons should not be taken as saying too much.

However, it also means that claims about specific higher-level vehicles (i.e., the precise set of neurons that answers to a specific content claim) rely largely on our content claims. Saying, of a neuron, that it is an “edge detector” is saying that it instantiates a higher-level vehicle (i.e., it is a set of one neuron) that can carry content about edges.
To adjudicate this claim, we need to know what it is for a vehicle to carry content (as well as how well this claim fits into an overall theory of neural processing). Before discussing content, however, let me address referents.

3.2.3. Referents

Referents are the external objects that representations assign properties to. One traditional way of determining the “reference” (a closely related, but distinct notion) of a representation is to rely on causes. However, as Dretske (1981, pp. 26–33) rightly notes, a theory of cause does not tell you which causes are important for representational content – which, in this case, means which causes are referents.

So the relevant causes need to be somehow “specially” related to the vehicles that are carrying content about them. That is, vehicles carry contents about things that they are, in some sense, good at carrying information about. So I need to say both what this special relation is, and what causes are, in order to determine what will count as the referent of a vehicle.

Of course, discussions of the nature of cause can be lengthy enough in themselves. Here I simply adopt David Fair’s (1979) approach, which identifies token causes with the transfer of energy momentum (“energy” from now on)\(^6\). For example, gas tanks cause gas gauges to register gas levels because there is a transfer of energy between gas tank levels and gas gauges.

More specific to semantic theories is the question of the “specialness” of the relation between the referents and the vehicles. Notably, both the referents, \(r\), and the vehicles, \(v\), can be measured in their own relevant units (i.e., the decoded alphabet and encoded alphabet, respectively). Many aspects of the relation between two such measurable quantities is determined by their joint probability distribution, \(p(r,v)\). This distribution can be used to determine to what degree the quantities are statistically dependent. If two quantities are dependent, then there is some relation or other between changes in one and changes in the other. A good way to quantify the degree of statistical dependence is to rely on a measure called “mutual information.” As a result, I use the two terms interchangeably\(^7\).

In fact, I think that mutual information is precisely the quantity we need to underwrite referent (and content) claims. Specifically, let me propose a preliminary version of the “statistical dependence hypothesis:”

The referent of a vehicle is the set of causes that has the highest statistical dependence with the vehicle under all stimulus conditions.

This hypothesis makes two important assertions: (1) the highest statistical dependency picks out referents; and (2) referents are causes, i.e., they have an energy transfer

---

\(^6\) Similar theories are suggested by Aronson (1971), Castaneda (1984), and Strawson (1987).

\(^7\) A different, but related, semantic theory that employs mutual information measures can be found in Usher (2001).
with the relevant vehicles. Cause plays a central role here as it rules out “accidental” statistical dependencies (e.g., the clock on the Peace Tower striking 12 and my going to lunch in Montreal). In general, statistical dependencies are too weak to properly underwrite a theory of content on their own.

Despite incorporating causes, this preliminary version of the hypothesis is still inadequate. In particular, it results in a kind of solipsism. This is because the highest dependency of any given vehicle is probably with another vehicle that transfers energy to it, not with something in the external world.

Fortunately, this is a rather artificial problem. Theories of representation are precisely theories that describe the relation between a representational system and what it represents. As a result, my previous identification of the system can be invoked here to limit the application of the statistical dependence hypothesis. That is, we only apply such a referent-determining rule external to the system. This results in a modified, and final, version of the statistical dependence hypothesis:

The referent of a vehicle is the set of causes that has the highest statistical dependence with the neural responses under all stimulus conditions and does not fall under the computational description.

The computational description here refers to the account of neural functioning provided by the theory of neural representation and computation discussed earlier. This description is adequate for only the behavior of neurons; that is, the components of the system.

A remaining concern with this formulation of the statistical dependence hypothesis pertains to the concept of “stimulus conditions.” There are a number of options for dealing with stimulus conditions, but let me just consider one here: appeal to other scientific theories to individuate stimulus conditions. Thus, stimulus conditions change if any physical variable relating the referent and the representation changes. Such variables include things like distance, illumination, relative velocity, etc. However, this definition makes it difficult to distinguish incremental changes in these more intuitive stimulus condition variables from incremental changes in variables that specify the physical properties of the referent itself. This is not a problem for the theory I have presented, although it may be a problem for determining what the referent is (i.e., in answering the question “is it a dog or a cat?”). Some such “stimulus conditions” would be difficult to realize because we cannot construct “dogcats.” That is, we cannot “morph” dogs into cats in the real world. This may limit our ability to systematically examine “all possible” stimulus conditions. Although this may initially seem problematic, the real-world development of an animal’s conceptual categories would also be so limited – so it may not be a hindrance after all.

3.2.4. Content

Recall that I take content to be the properties ascribed to a referent by a vehicle. Recall as well that the theory of neural representation outlined earlier can be expected to
play a central role in content determination. However, as discussed, it does not by itself tell us how to make content ascriptions. This is because it is strictly a theoretical or computational description of the functioning of neural systems. In order to inform the theory of content, this computational description needs to be related to a semantic one.

Recall that a neural representation is defined by identifying an encoding and decoding process. Previously, I mentioned that there are methods for finding the representational decoder that allow very good reproduction of the encoded signal. In other words, the decoder tells us how to relate the neural signal to the encoded signal, which means it tells us what properties of the encoded signal are “saved” by the neural signal. This in turn tells us which properties are ascribed to a referent by neural activity.

So content is determined by the decoders. But how are the decoders determined? Simply put, they are found by minimizing the difference (an “error” or “energy”) between the signals being represented and the decoded neural activities over all represented input signals. This tells us that the decoders depend on the relation between two things: (1) the signals to be represented, s; and (2) the neural activities, or response, r. The response, r, is the activity of a vehicle, and the signals, s, are the referent.

In effect, then, the decoders describe a rule that determines what properties the current neural activities in a population of neurons ascribe to the current referent of the population. That rule is determined by examining the statistical dependence the vehicle has with a referent over all stimulus conditions.

This description of content highlights an important distinction between two kinds of content, what I will call “occurrent” content and “conceptual” content. The former applies to current neural activity and the current referent, under the current stimulus conditions. The latter applies to the determination of the decoders, over all stimulus conditions. Noticing this difference makes it clear that the statistical dependence hypothesis is applicable for the determination of conceptual content. To extend this application to occurrent content, let me suggest the following corollary:

The occurrent referent of a vehicle is the set of causes that has the highest statistical dependence with the neural responses under the stimulus conditions in which it occurs.

Thus, the process of content determination stays the same, but referent determination varies across these two kinds of content. I will revisit this distinction in more detail in my account of misrepresentation. For now, it merely needs to be clear that the statistical dependence that holds over all stimulus conditions can be quite different from that that holds under a particular subset of those stimulus conditions. According to this theory, such a difference in statistical dependence results in a difference in referent. The properties ascribed (i.e., occurrent content) to such a referent by neural activity are determined by the decoders, which are themselves determined by property ascriptions under all stimulus conditions (conceptual content).

8 A technical discussion of how these are combined can be found in Eliasmith and Anderson (2003).
3.3. Discussion

Because statistical dependence plays such an important role in this theory, it is worthwhile discussing it in more detail. Adopting statistical dependence as a means of referent determination has a number of beneficial consequences. First, statistical dependence comes in degrees. The highest dependence of one vehicle with its referent may be higher than the highest dependence of another vehicle with its referent. The strength of the dependence maps nicely onto the precision of the representation in question. If, for example, the occurrent dependence of my representation of a dog’s position with the dog’s actual position is nearly perfect (which means any changes in my relation to the dog are reflected in changes in the contents of my vehicle), we know that my representation is precise.

A second benefit of using statistical dependence is that it can help minimize our reliance on intuitions about what is represented by cognitive systems. In other words, it provides a means of systematically examining the features of the environment on which the vehicle is statistically dependent; we will be in a position to discover, not stipulate, what the referent of the vehicle is. In other words, it helps relieve us of our linguistic bias, which has informed most past theories of categorization.

If we have a set of neurons that we believe to be a vehicle, we can explicitly construct the joint probability histogram between those neurons and features of the environment and thus discover the referent of that vehicle. In other words, we can test our hypotheses about possible vehicles. If we suppose that a certain set of neurons acts as a “dog” vehicle, we can test that hypothesis by seeing how statistically dependent the vehicle and referent are. If they have no dependencies, or have better dependencies with other vehicles or referents, our hypothesis about the nature of the vehicle will change. The converse may occur as well: what we take to be good referents will help us determine vehicles. These methodologies are complementary and serve to give us a good idea of what vehicles and referents there are. What is most important is that this theory allows us to make discoveries; no doubt we will be surprised by some of the things biological systems depend on for representing the environment.

The last consequence of adopting statistical dependence I will discuss is a subtler one. Because referents are related to higher-level vehicles by statistical dependencies, our identification of referents and of higher-level vehicles is mutually dependent. Furthermore, referents help determine contents, so contents depend on referents as well. Therefore, content depends on properties of vehicles. This should not be too surprising. If you know all there is to know about a vehicle, you will know its possible content. This is not to say that the vehicles we discover will determine content; rather, they will constrain possible content. Psychophysics, for example, tells us that there are certain dynamic ranges over which our retinal cells can encode brightness. Outside of those ranges, differences in intensity are indistinguishable; that is a fact about physiology. If the vehicles cannot carry such differences in intensity, then those differences cannot be used by the system to react to the environment; consequently, those vehicles cannot carry content about those differences.
What this means, then, is that vehicles and contents are not independent. In a more traditional turn of phrase: syntax and semantics are not independent\(^9\). The stuff that carries meaning (syntax/vehicles) also helps to determine meaning (semantics/content). As powerful as natural languages are, they are stuck with certain vehicles. There are contents that these vehicles carry well and there are contents they do not carry well. A fixation on language for understanding categorization is likely to bias our consideration of possible categories, potentially blocking the route to a successful theory.

4. Misrepresentation

Dretske (1995) notes that the problem of misrepresentation is above all the problem of explaining how we assign the wrong properties to an object [Dretske (1994) p. 472, (1995)]. This is the same as what Fodor (1987) calls the “disjunction problem.” It is the problem of explaining how a representation can assign a specific set of properties to a referent even though the representation can be caused by a disjunction of referents that do not all have those properties. For example, according to a naive causal theory if my “dog” vehicle is caused by a cat under some circumstances, that should mean that the vehicle is actually a “dog or cat” vehicle (since it can be caused by either). So the problem is how do I explain ascribing the property “dog” to a cat in certain cases (rather than just tokening a disjunctive vehicle)? How, in other words, can I explain when my categorizing a cat as a dog is a mistake. In this section, I address this problem and show how the theory I have presented can avoid it.

Philosophers tend to presume that cases of representing fall into one of two categories: right or wrong. This, it seems to me, is a mistake. Dretske (1994 p. 472), for example, defines misrepresentation as saying of something that does not have a given property, that it has that property; for example, saying that a black dog is brown. In other words, misrepresentation is representing one thing (a black dog) as another (a brown dog). Unfortunately, under a strict application of this definition, it is not clear that we ever get anything right – that we ever represent.

Consider my representing a black dog standing in front of me. Suppose that after 3 min the dog is removed from my sight. I would then be able to answer all sorts of questions about its shape, size, length, etc., based solely on my representation of the dog. However, each of my answers would be inaccurate in some way. Consider the line of questioning: What color was the dog? Answer: Black. Dark or light black? Answer: Dark black. This color (showing a color swatch)? Or this color (another swatch)? etc. There is little doubt that I would eventually answer incorrectly. Does that mean that I am attributing to this dog a property it does not have (i.e., a certain shade of black)? Yes, it does. Does it mean that we should say I am misrepresenting the dog? According to

\(^9\) For arguments concerning the benefit of blurring the distinctions between vehicles and contents, see, e.g., Langacker (1987) and Mohana and Wee (1999). See Eliasmith and Thagard (2001) for an example of how blurring the distinction may help explain the nature of high-level cognitive processes.
Dretske’s definition it does, but I think that such an answer is overly hasty. In order to say why, consider some elementary distinctions in measurement theory.

Measurements are said to be accurate if they are near the right value. If I measure darkness and get the right answer, I have made an accurate measurement. Measurements are said to be precise if they are reproducible. If I measure the darkness of a color swatch over and over again, and get the same answer every time, I am making precise measurements of darkness. If my measurements are precise and accurate, they are said to be exact. Notably, precision is a property of a set of measurements, while accuracy is a property of a single measurement. But we can define the accuracy of a set of measurements as the average nearness to the right value. In statistical terms, precision is measured by the variance of a set of measurements, while accuracy is the difference between the average measurement and the correct answer.

What the line of questioning above is doing, is probing the representer with increasing degrees of precision. Although the representer may be perfectly accurate at one degree of precision (black versus white) the representer may be inaccurate at another (one color swatch versus another). Neural representations have a limited degree of precision; only about three bits of information are transmitted per spike [Rieke et al. (1997)]. So we should not be surprised that we will eventually misascribe properties; it is not possible to be consistently accurate to a degree of precision greater than our “measuring device” can provide.

What is important here is that representations are best characterized as “better” or “worse,” not “right” or “wrong.” “Better” means high accuracy with high degrees of precision. “Worse” means low accuracy with low degrees of precision. We may want to make “right” and “wrong” claims at a given level of precision, but in doing so we would have to make an argument as to why being below some standard of accuracy at a given precision is a good criterion for making this distinction. Dretske’s definition clearly does nothing of the sort. Using the term “misrepresentation” to divide representations into two groups obscures important subtleties of representation. We will have a more general understanding of misrepresentation (i.e., one that does not depend on choosing particular standards and can account for degrees of deviation from any given standard), if we accept that representations come in degrees, i.e., that they lie on a continuum from good to bad.

The preceding discussion lays the groundwork for addressing Fodor’s disjunction problem head on. The challenge of the disjunction problem, then, is to explain how my representation of something (a cat) could mean to me that it had some property (the property of being a dog) even though it was caused by something that I would say did not have that property (a cat). The solution is that we can explain this case of misrepresentation by a careful application of the statistical dependence hypothesis. More specifically, we can notice that the hypothesis and its corollary give different answers. In particular, under all stimulus conditions, this vehicle has the highest statistical dependency with dogs even though, under this condition, this vehicle has the highest dependency with a cat.

But, does this really solve the disjunction problem? Will it not be the case that the highest statistical dependency holds between dogs-or-this-cat under all stimulus conditions?
In fact, no. This cat under all stimulus conditions will not have a high statistical depend-
ency with my “dog” vehicle; it will have the highest dependency with my “cat” vehicle. It is only under this stimulus condition that it has a high statistical dependency with my “dog” vehicle. In other words, because there is another vehicle (the “cat” vehicle) that has a higher dependency with this referent under all stimulus conditions, it cannot be this vehicle (the “dog” vehicle) that has this cat as its referent. Note also that this solution is possible because the notion of representation/misrepresentation is a graded one.

We can introduce a further variant in order to try to preserve the disjunction prob-
lem. That is, perhaps the problem is that my vehicle has the highest statistical depend-
ence with dogs-or-this-cat-under-these-conditions. Luckily, this disjunction can be ruled out because it includes a specification of stimulus conditions. We cannot find a dependence between a vehicle and something-under-a-stimulus-condition under all stimulus conditions since most of the stimulus conditions are ruled out by such a char-
acterization. It would be self-contradictory to try and determine such a dependency. Thus, I take it that the account I have given of content determination can satisfactorily account for misrepresentation.

5. Conclusion

The theory of mental meaning I have presented begins with considerations of represen-
tation in neural systems. For this reason, I have called it a theory of neurosemantics. However, because I have identified explicit relations between neural and higher-level representations, I take the theory to be applicable to the more traditional problems of psychosemantics. I have shown how the referent relation, and the statistical dependence hypothesis on which it relies, can be used to address at least one such problem, that of misrepresentation. There are, of course, many issues which I have not addressed in this paper, but I take this kind of initial success to bode well for neurosemantics in general.

Finally, the fact that I have successfully adopted a neuroscientific view of the problem of naturalizing semantics demonstrates that, indeed, the “neuron-by-neuron” account so scorned by Lycan and others may turn out to be the most fruitful one. The irony that try-
ing to understand categorization often requires us to divest ourselves of our own cherished categories can best be embraced by the adoption of quantitative tools applied to charac-
terizations of the system we are trying to understand. At the moment, neuroscience (com-
bined with engineering), more than linguistics, philosophy, or psychology, provides the kinds of descriptions of cognitive systems that can inform such a characterization. Getting to the heart of categorization is a matter of properly understanding the biological imple-
mentation of cognitive systems: once again, the devil is in the details.

References

Fodor, J. (1975), The Language of Thought (Crowell, New York).
Goodman, N. (1968), Languages of Art, (Hackett Publishing Company, Indianapolis, IN).
Chapter 48

CONCEPTUAL ANALYSIS AND PHILOSOPHICAL NATURALISM*

ELISABETTA LALUMERA

Università degli Studi di Bologna, Italy

Contents

Abstract 1055
1. Introduction 1056
2. What is intuitive about conceptual analysis? 1057
4. Against privileges 1059
5. The inward approach 1061
6. Conceptual truths or truths about concepts? 1062
7. The outward approach 1064
8. ‘Bachelors are unmarried men’ is about facts 1065
9. Explaining away the illusion 1067
10. A “mixed bag” 1068
11. Conclusion 1070
References 1070

Abstract

This chapter focuses on a recent revival of conceptual analysis, the philosophical method of discovering necessary and a priori contents by describing conceptual relations. According to some philosophers, contents such as ‘Red is a color’ are true by virtue of the deep structure of our cognitive system, and assuming that this structure is innate, such contents are true independently of experience, i.e., they are a priori. I defend an alternative position, according to which relations among concepts, whether innate or acquired, mirror the relations among the real-world properties they refer to, and our conceptual structure is continuously tested and compared with the characteristics of our environment. In such a view, if a description of our conceptual equipment generates true contents, it generates a posteriori ones.

*I am grateful to Claire Lefebvre and Henri Cohen for their invitation and support, and to the audience of my talk at UQAM for helpful suggestions. My thanks are also due to Paolo Leonardi for discussions of most of the ideas contained here.

Handbook of Categorization in Cognitive Science, Edited by Henri Cohen and Claire Lefebvre
© 2005 Elsevier Ltd., All rights reserved
1. Introduction

Many philosophers have recently proposed a revival of conceptual analysis, or the method of discovering necessary and *a priori* contents by describing conceptual relations. Supporters of conceptual analysis start from the fact that people intuitively tend to attribute a special epistemic status to some contents, which seem to be true come what may, in all possible circumstances, and solely by virtue of the nature of their components. Familiar examples of such contents are definition-type contents such as ‘Bachelors are unmarried’ and ‘Vixens are female foxes,’ and categorical judgments like ‘Schnauzers are dogs’ and ‘Red is a color.’ Traditionally, philosophers set themselves the task of finding out more interesting ones, involving, say, our concepts of good, virtue, knowledge, meaning, and concept itself.

In this chapter, I will concentrate on the prospects of reviving conceptual analysis from within philosophical naturalism (in Quine’s sense). I will compare two different attitudes that a naturalistic philosopher may take toward the intuition of special-status contents. At the moment, two kinds of answers have been proposed. I will call them the *inward approach* and the *outward approach*. The inward approach combines the claim that whatever is true can be discovered, at least in principle, by natural science, with an enthusiasm for conceptual analysis. According to the inward approach, sentences like *Nothing is red and green all over* are true by virtue of the deep structure of our cognitive system—most likely of our language faculty. Assuming that this structure is innate, special-status contents are true independently of experience, and therefore they are *a priori*. In this view, “cognitive psychology” is the new name for conceptual analysis. Rey has developed this idea in some recent papers [Rey (1998, 2001, 2004)], drawing on Chomskian materials; similar positions are held by Peacocke (1998a,b), and Horwich (2000).

The outward approach to special-status contents is typical of externalist theories of mental content—for example, it can be found in Fodor’s (1976, 1998) causal semantics and in Millikan’s (1984, 1998, 2000) teleological semantics for concepts. According to the outward approach, concepts are generally individuated in terms of what they are the concepts of. Even for contents like ‘Bachelors are unmarried men,’ there is some fact of the matter that makes them true, which can be discovered *a posteriori*. According to this approach, no content is true solely by virtue of the nature of its components—in other words, there are no special-status contents of the kind envisaged by the tradition of conceptual analysis. Nevertheless, a supporter of the outward approach needs to explain why some contents *seem* to have a special status.

In what follows, I will assess the inward and the outward approach as candidate explanations of intuitions of special-status contents. My conclusion will be that the inward approach suffers from a structural dilemma, which may undermine its prospects of reviving the tradition of conceptual analysis from within a naturalistic perspective. The outward approach, on the contrary, is revealed as a coherent option, although it faces a conspicuous challenge.

The chapter is organized as follows. In Section 2, I will discuss what I take to be the intuition behind conceptual analysis. Then in Sections 3 and 4, I will sketch the picture
behind the tradition of conceptual analysis, and summarize the main problems that have been raised for this approach in the last few decades. Sections 5–10 examine the inward approach and the outward approach, and Section 11 presents my conclusion.

2. What is intuitive about conceptual analysis?

The project of conceptual analysis appears to have an intuitive core. The intuition is that some contents, such as ‘Vixens are female foxes,’ ‘Two is a natural number,’ ‘Nothing is red and green all over,’ and ‘Schnauzers are dogs’ seem to be true come what may. You will never meet a vixen that is not a female fox, however far you might go. You can be sure that even if you missed a couple of copies of Scientific American, no article about an entirely red and green thing was in there. And if someone tells you that her father is a schnauzer, you will not believe her. Compare a second group of contents containing the same concepts: ‘Vixens are vegetarian,’ ‘Two is the number of the dogs I own,’ ‘Nothing is entirely red and a national flag,’ ‘Schnauzers are color-blind.’ Before believing the contents in the second group to be true, you will look for evidence. You may also try to falsify them, to find counterexamples. None of these measures seem to be appropriate with contents of the first kind. Why should this be so?

Here intuition ends and the classical philosophical answer begins: “Because the contents in the first group, unlike those in the second, and the sentences expressing them, express what is in our concepts [VIXEN], [TWO], [RED] (or [GREEN]), and [SCHNAUZER], respectively. That is why they are true come what may. And that is why they have to be true. By analyzing concepts, that is, by specifying all their proper parts in a definition, one will find more of such privileged contents. Some analyses will be easy, as in the case of [VIXEN]. Others will not – think of concepts like [GOOD], [KNOWLEDGE], and [MEANING]. But it is just a question of refining the method.”

Note that the point here is not that, in most cases, people would assent to ‘Vixens are female foxes’ more quickly than they do to ‘Vixens are vegetarian’ – that some category judgments are easier to produce than others, in quantitative terms. Rather, the idea is that, upon reflection, people tend to consider some judgments containing a certain concept as having a different status from others, and they generally agree on which these judgments are. This was already Grice and Strawson’s (1956) point against Quine’s “Two dogmas” [Quine (1951)]. In the same vein, Katz (1972) argues that people have semantic intuitions in addition to syntactic intuitions – in both cases, these can be taken to be regularities in judgments that we need to explain. Just as speakers of English tend to agree on the judgment that

* To who is the man I gave my book?

is syntactically incorrect, they would also agree on judgments of synonymy, that is, they would differentiate contents like ‘Vixens are female foxes’ from contents like ‘Vixens are vegetarian.’ If Katz is right, then theories of concepts, as well as theories of language knowledge and acquisition, have data to start with. The problem is what this intuition, or these data, are supposed to mean.
3. Cognitive privileges, metaphysical privileges, and the Transparency Thesis

One interpretation is provided by the classical philosophical answer I mentioned above. It may be described, less colloquially, as consisting of six interrelated claims. I am not suggesting that such claims correspond to the view of any particular philosopher; in fact, it may well be the case that no one has explicitly subscribed to all of them. But they may be taken as defining an ideal picture, which may be useful as a term of comparison and as a starting point for discussion. Here is the ideal picture:

1. At least some concepts are complex, and they have parts that are themselves concepts.
2. The relations a concept bears to its components are constitutive of its identity: they make it the concept it is.
3. Some knowledge of the relations between a complex concept and its components is constitutive of people’s possession of that concept.
4. Sentences and thoughts expressing constitutive conceptual relations are analytic, that is, they are true by virtue of what the concepts and terms involved mean.
5. Sentences and thoughts expressing constitutive conceptual relations are a priori, i.e., justifiable independently of experience.
6. Sentences and thoughts expressing constitutive conceptual relations are necessary.

The ideal picture explains semantic data with a two-tiered theory. Claims 1–3 express a theory of concepts. They say that some of the relations a concepts bears to other concepts are cognitively privileged. Each (complex) concept bears intrinsic relations to its components, and such relations account for its identity; moreover, having such relations represented in one’s mind accounts for one’s possession of that concept. This is (roughly) the definitional view of concepts, or the Classical Theory. According to the Classical Theory, complex concepts are equivalent to sets of simpler components. Correspondingly, to possess a complex concept is to have such components represented in one’s mind.

Conversely, claims 4–6 express a metaphysical theory. They say – roughly – that the structure of our way of representing the world corresponds to how the world is. We read off necessary truths from our concepts. According to claim 5, for example, we can learn that there are no entirely green things that are entirely red at the same time simply by checking the definitions of [RED] and [GREEN]. And according to claim 6, we learn that there cannot be things that are entirely red and green at the same time. I will call the idea that truths about the world can be read off from conceptual relations the “Transparency Thesis,” for it implies that by looking at our concepts, we actually see how things are in the world. With the Transparency Thesis, claims 4–6 say that the very same contents that are cognitively privileged are also metaphysically privileged.

The Transparency Thesis has always been considered a powerful tool to employ in philosophical arguments. It makes it possible to reject an opponent’s claim on the basis of the definitions of the concepts of which it is composed. This partially explains why philosophers have always cared so much about conceptual analysis. A traditional
application of the Transparency Thesis is to antiskeptical arguments. Fodor (1998, p. 70) sketches the typical dialectic very effectively:

Sceptic: You can’t ever infer with certainty from how things look to how they are.
Antisceptic: Can too, because there is an intrinsic conceptual connection between how-things-look concepts and how-things-are concepts (between behavior-concepts and mind-concepts; between is-concepts and ought concepts, etc. etc.). Bop. I win.
Sceptic: I don’t acknowledge such intrinsic connections.
Antisceptic: Then you don’t have the concepts! Bop. I still win.

But how can it be that, by looking at how concepts are structured, we can also see how the world is structured? How can the Transparency Thesis be supported? Different philosophers over the centuries have employed different strategies. Kant, for example, relied on nothing less than the transcendental deduction. Moore thought that the constituents of mental and linguistic contents, i.e., concepts, were one and the same thing as the constituents of reality. Later twentieth-century versions of the conceptual analysis program depended on two well-known moves. The first is the analytic explanation of apriority: contents that are true by virtue of the meaning of their terms can be justified (and believed) independently of experience. Many philosophers adopt a stronger version: contents that are true by virtue of the meaning of their terms cannot be falsified by experience.\(^1\)

The second move is the identification of apriority with necessity: whatever we can know independently of experience could not have been otherwise. In particular, given that analytic contents are \textit{a priori}, contents expressing constitutive conceptual relations are also necessary truths. So, for example, on the hypothesis that ‘Schnauzers are dogs’ is true simply by virtue of the structure of the concept [SCHNAUZER], we cannot categorize something as a schnauzer without also categorizing it as a dog. It would then follow that it is inconceivable that schnauzers are not dogs, and therefore, according to this view, it would be necessary for schnauzers to be dogs. With both moves in place – the analytic explanation of apriority and the identification of apriority with necessity – the Transparency Thesis is secured: by looking at our concepts, we see through what there is and what there has to be. Conceptual analysis is not just the analysis of concepts; rather, it is the analysis of what is the case.

4. Against privileges

As is well known, in the last half-century the ideal picture has been attacked from different sides. From one side, the cognitive privileges described by the Classical Theory were questioned. In particular, the Classical Theory of concepts was shown to be unable

\(^1\) The distinction between a stronger and a weaker sense of \textit{a priori} is presented in Kripke (1981, p. 35). The stronger sense was presupposed in Quine’s (1951) attack to the notion.
to explain the phenomena of typicality and preferred levels of categorization identified by Rosch [(1973), Rosch et al. (1976)]. The first problem with the Classical Theory is that it admits of no gradation – anything that satisfies the definition associated with a certain concept ipso facto falls under that concept. There is no room for explaining that some concept applications are easier to make, learned earlier, and more frequently employed. Moreover, it is a fact that only a few examples of conceptual definitions have been provided so far, and those only trivial ones, such as ‘Bachelors are unmarried men.’ Some have argued that this fact must be a symptom of the psychological implausibility of the Classical Theory [Fodor (1998)]. The same conclusion is supported by Wittgenstein’s (1953) suggestion, supported by psychologists, that we categorize things as belonging the same concept even when there is no set of properties they all share, as long as we can identify some family resemblance, or nontransitive aspects of similarity, among them. For all these reasons, it is now widely believed that, as a theory of what people have in mind when they categorize reality, the Classical Theory of concepts cannot be the whole story; there are no clear-cut cognitive privileges of the sort many philosophers envisaged.

From another side, the ideal picture behind the program of conceptual analysis was the target of Quine’s (1951) attack on the analytic explanation of apriority, and of Kripke’s (1971) objections to the identification of apriority and necessity. Quine’s “Two dogmas of empiricism” is generally read as a definitive demolition of the idea that something can be true solely by virtue of meaning. Through an argument for exclusion, Quine rejects all alleged accounts of analyticity (in terms of cognitive synonymy, explicit convention, or modal properties of sentences). He also puts forward the idea that no content is immune to empirical disconfirmation, not even alleged definitions – sentences “confront the tribunal of experience only as a corporate body” (p. 42) – thereby rejecting the stronger version of apriority. Though some of Quine’s original arguments may be shown to rely on his behavioristic convictions, his conclusion has been accepted; after Quine, no one really attempted a meaning-based account of apriority.

As for the link between apriority and necessity, conclusive examples have been given of a priori contents that are not necessary, and of necessary contents that are not a priori, thereby invalidating the a priori explanation of necessity. ‘I am here now,’ proposed by Kaplan (1989), can be known a priori without being necessary; Kripke’s ‘Water is H₂O’ is arguably necessary but discovered a posteriori.

The point of my long digression on the abolition of privileges was to point out a gap that has opened up between the two parts of the ideal picture. Without the analytic explanation of the a priori and the identification of apriority with necessity, the Transparency

---

2 In fact, philosophers who engaged in conceptual analysis, from Descartes to Kant, from Frege to Wittgenstein, Austin, etc., did not take concepts to be what people have in mind when they categorize. Rather, they took concepts to be abstract objects to which we are epistemically related, either by intuition or through language understanding. Therefore, strictly speaking, the psychological implausibility of definitions does not affect the ideal picture, on the hypothesis that concepts are abstract, nonpsychological entities. It does, however, affect the idea of employing the ideal picture as a psychological theory of concepts, or better, the idea of fitting a psychological theory of concepts into the ideal picture.
Thesis stands in need of supporting arguments. But the Transparency Thesis was precisely what licensed the identification of cognitive privileges with metaphysical privileges within the ideal picture. Once the Transparency Thesis is discarded, one is no longer warranted to believe that truths about the world can be read off from our conceptual structure, or that metaphysical privileges correspond to cognitive privileges. At least, we cannot believe this without further arguments.

5. The inward approach

The inward approach inherits most of the claims of the ideal picture underlying the conceptual analysis of the past:
1. At least some concepts are complex, and they have parts that are themselves concepts.
2. The relations a concept bears to its components are constitutive of its identity: they make it the concept it is.
3. Some knowledge of the relations between a complex concept and its components is constitutive of people’s possession of that concept.
4. Sentences and thoughts expressing constitutive conceptual relations are a priori, i.e., justifiable independently of experience.

Proponents of the inward approach usually generalize from the hypothesis that some form of basic logical competence is innate and deep-level to the more ambitious claim that meaning postulates or conceptual definitions are innate. Here is how Horwich (2000, p. 168) explains the hypothesis of in-born logic:

Suppose that each human being is born with, and stuck with, a simple language of thought (i.e. mentalese) containing, amongst other things, certain symbols whose intrinsic nature is such that the principles of classical logic are obeyed. In that case we would have an a priori commitment to classical logic. For this commitment would neither derive from experience nor be revisable in light of it.

If we can imagine that a deep inquiry into our conceptual competence – roughly, into the semantic functions of our linguistic faculty – would reveal that we are loaded with logical principles, we might also imagine that we are equipped with conceptual definitions. These deeply buried but nevertheless psychologically real conceptual definitions would explain our intuitions concerning privileged contents – in fact, privileged contents would be a symptom of the existence of such deep rules. In other words, just as syntactic data are a clue for the linguist to discover the deep rules of our knowledge of syntax, semantic data would represent a clue for the cognitive scientist to discover the inner rules of our conceptual competence. For example, the fact that we find ‘Bachelors are unmarried men’ to be privileged would reveal that the concepts [NOT MARRIED] and [MAN] are parts of the concept [BACHELOR]; the specific fact that we find such content to be true come what may would suggest that we know that bachelors are unmarried men not by experience, but by virtue of the innate deep structure of our mind [Rey (2001, 2004)].
Here, innateness plays the role of the source of a priori knowledge: if something is known, so to speak, from the start, and inferred as the output of one’s linguistic faculty, then it is not known by experience; therefore, it is knowable independently of experience; and therefore, it is a priori.

Note that the strategy of locating cognitive privileges in the deep structure of our competence circumvents the familiar objection based on the unavailability of definitions. As is well known from the debate concerning the Classical Theory of concepts, there are only a few definitions that we can produce or even recognize. So if definitions are psychologically real, they have to be beyond the range of introspection; they must operate somehow down there\(^3\). In this sense, conceptual analysis would be the same as cognitive psychology.

A supporter of the inward approach may even turn the very fact that we can neither produce nor recognize more than a few definitions into a case for her position. The argument would go as follows. It is a fact that we cannot provide definitions for most of our everyday concepts, such as [CHAIR] or [COLOR]. But surely it is also a fact that we are competent in the use of such concepts. Therefore, either concepts are not definitions, or concepts are definitions that are below the level of introspection. On the hypothesis that concepts are definitions, conceptual definitions have to be deep-level, as the inward approach would have it. In fact, this is a sort of transcendental argument – given the facts about our performance, our competence must have such and such a nature.

This form of argument is explicitly in Peacocke’s writings. As a reinforcement of his dialectic, he proposes the following case. Leibniz and Newton both employed the mathematical concept of limit of a series, but they never succeeded in providing a successful definition; however,

[j]It would be huge injustice to Leibniz and Newton to deny that they had the concept of a limit of a series, or to deny that they had propositional attitudes describable by using the word “limit” within the “that” clauses... I would say that each of these great thinkers had an implicit conception which guided their application of the phrase “limit of...” in particular applications.

Peacocke (1998a, p. 49)

6. Conceptual truths or truths about concepts?

In the ideal picture of old-style conceptual analysis, the quest for definitions, or Lockean nominal essences, was one and the same thing as the quest for the nature of things, or real essences. Conceptual privileges were expressed by contents that were a priori true, and metaphysically necessary. The link between cognitive privileges and metaphysical privileges, as we saw above, was secured by the Transparency Thesis, recently supported, in its turn, by the analytic explanation of the a priori, and by the linguistic theory of necessity.

\(^3\) See Peacocke (1998a,b) for the idea that concepts are defined by implicit conceptions.
Now, nothing of this sort is available to the supporter of the inward approach. When she claims that ‘Bachelors are unmarried’ and ‘Nothing is red and green all over’ express privileged relations among concepts, she means that such sentences are solely about the structure of our concepts. When we say that nothing is green and red all over, it is concepts, and not colors, that we are talking about. So, if we adopt the inward approach, we might well know *a priori* that the concept [GREEN] and the concept [RED] cannot be applied to the same thing at the same time. But what about the colors red and green: can they coexist on the same surface at the same time? Likewise, it is all very well that the concept [SCHNAUZER] has the concept [ANIMAL] as a proper part, but no conclusion about schnauzers (the dogs) can thereby be inferred.

Why such a limited achievement? Because there is no Transparency Thesis to license the passage from cognitive privileges to metaphysical privileges; there is no guarantee that our conceptual structure mirrors the structure of reality. Just as Chomskian rules are about the inner functioning of a mechanism, the language faculty, so the conceptual rules advocated by the inward approach “cut no ice” – they describe the inner functioning of another mechanism. Accordingly, Rey (1998, p. 39) remarks that genetic-based conceptual truths are not necessary truths about the world.

Rey’s remark is surely correct. But then a more general problem arises concerning the status of conceptual rules. Even on the milder hypothesis that they do not generate necessary truths, the supporter of the inward approach still claims that they generate truths. It is true, for example, that if the concept A contains the concepts B and C, then all Bs are As. Now the question is how to *validate* conceptual rules from the point of view of the inward approach. How are we entitled to claim that our rules – say, the rule that connects the concept of [GREEN] to the concept of [RED] – are good, truth-preserving rules? To frame it more skeptically, nothing so far excludes the possibility that we are born with bad conceptual rules, involving concepts that, like [PHLOGISTON], [WITCH], or [GRUE], individuate no genuine property, and *a fortiori* no necessary relation between properties.

A simple explanation is at hand: *experience* (of the species, or of the individual, when possible) sorts out good (truth-preserving, useful) conceptual rules from spurious ones. For example, one could say that contents generated by the [PHLOGISTON] rules have not proved useful or explanatorily valid when tested. But a supporter of the inward approach cannot avail himself of the idea of validation through *experience*, for that would conflate with his main concern, that is, giving a new foundation to the *a priori*. The only option is to resort to a weak conception of the *a priori* whereby being *a priori* is compatible with being disconfirmed by experience.

It therefore seems that the inward approaches force us into a dilemma. On the one hand, there is the idea that intuitively privileged contents are just about the structure of our mind. In this case, contents expressing cognitive privileges would be true, *a priori*, and maybe also necessary, but would cut no ice in Millikan’s sense. What we know *a priori* would be not that vixens are female foxes, but rather that our concept [VIXEN] contains the concepts [FEMALE] and [FOX]. (Try other examples to test what this amounts to.) This particular horn of the dilemma sounds coherent, but it would undermine what I
think is an even more plausible intuition, that is, the intuition that we rarely talk about concepts; instead, most of the time we talk about real-world properties and facts. Anyway, we certainly seem to be doing so in the case of ‘Red is a color,’ ‘Schnauzers are dogs,’ and even ‘Vixens are female foxes.’

On the other hand, there is the temptation to claim that cognitive privileges correspond to metaphysical privileges, so that ‘Vixens are female’ is not just a conceptual truth, but also a truth about the world that we get to know via concept analysis. But here, something more from the inward approach is needed, in the form of a surrogate of the Transparency Thesis, that is, an explanation of how, by inspecting the conceptual structure, we might be able to read off the structure of reality, either in its necessary features, or at least in its genuine ones.

7. The outward approach

What I have called the outward approach to the intuitions of privileged conceptual relations bites the bullet and acknowledges that the Transparency Thesis is no longer sustainable. Causal theories and teleological theories of concepts may be described as adopting such an approach. What is common to such theories is the view that concepts are individuated in terms of what they are the concepts of, and not in terms of their relations to other mental contents. So, for example, what makes the concept [Schnauzer] the concept it is is not its relation to my [DOG] concept or to my [PET] concept, but rather the fact that this concept is my mental indicator and detector of a certain species of dogs, that is, schnauzers. It is the concept that originated from my causal contact with schnauzers, and that I use for keeping track of and representing schnauzers. The causal and/or teleological connection with a certain aspect of the world is what establishes the identity of the concept – in general, external mind-world relations are in charge of concept identity [Fodor (1998), Millikan (2000)].

This does not amount to saying that externalist theories deny that people associate images, descriptions, and information in general with each of their concepts. Rather, they would acknowledge that such information is usefully projected from one concept application to the others. For example, if I associate the piece of information (likes chewing bones) with my concept [SCHNAUZER], I can predict that the next thing I categorize as a schnauzer will like chewing bones. But no piece of information is necessary in order for me to have the concept. A corollary of this point is that two different individuals can both have the concept [SCHNAUZER], even if there is no piece of information, definition, or image they both possess, as long as they both have a concept of schnauzers.

According to the outward approach, there are no conceptual definitions that would establish the conditions of concept identity and concept possession; therefore, there are no cognitive privileges of the sort that the Classical Theory of concepts envisaged. So intuitively privileged contents (Katz’s and Rey’s “semantic data”) are not about part-whole relations among concepts. Like any other contents, they are about concepts’ referents, i.e., properties. Thus, the outward approach captures the intuition that what
we talk about when we say that schnauzers are dogs or that nothing can be both red and green, it is dogs and colors we are talking about, and not concepts. Note that given the dismissal of the Transparency Thesis, if these contents were about concepts, they would be only about concepts – they would “cut no ice”⁴, in Millikan’s words, as in general, the inference from conceptual-structure claims to world-structure claims is not valid.

As an approach to semantic data, the outward approach faces two main challenges. The first is to specify the factual truth ground of every special-status content. If a special-status content is not true by virtue of our conceptual structure alone, then one has to provide an alternative account of what its truth-makers are. This challenge is relatively easy to meet for some contents, like the above-mentioned ones about dogs and colors, because an ontological commitment to the existence of colors and dogs is not too demanding to our common-sense ontological convictions. But there are hard cases as well.

The second challenge for the outward approach is to explain why special-status contents strike us as special, even though all contents are on a par insofar as they are factual. In other words, we need to explain why there are semantic data, or contents, that people tend to view as special. I will deal with the first challenge first.

8. ‘Bachelors are unmarried men’ is about facts

Let us concentrate on the nastiest kind of case for the outward approach. According to the outward approach, ‘Bachelors are unmarried men’ is not about concepts; rather, it is about real-world properties. It is a fact (an arrangement of properties) that bachelors are unmarried men. But which properties are involved here? What is the factual component that would make such content true?

Fodor (1998) claims that ‘Bachelors are unmarried men’ expresses a metaphysically necessary relation holding between being a bachelor, being a man, and being unmarried. The latter properties would constitute the former, just as H₂O constitutes being water.

Now, this solution is viable if and only if a plausible ontology of properties – i.e., concepts’ referents – is provided. In the case of ‘bachelors are unmarried men,’ the challenge for the outward approach is precisely that of making sense of adding a property like being a bachelor to the inventory of what exists, when the component properties (being a man, being adult, etc.) are already present. According to some philosophers, this would already count as an objection to Fodor’s version of the outward approach: no such luxurious ontology is defensible, and therefore the outward approach fails. A more sympathetic approach, which I favor, would be to maintain that it is an open question whether the task can be accomplished. In fact, some work has already been done in this regard⁵.

---

⁴ Millikan [(1984), p. 272] holds the radical view that logical principles are grounded in natural necessities.
⁵ See Fodor (1998) for mind-dependent properties and Millikan (2000) for an extension of the class of natural kinds to culturally, socially, or historically based groupings of things.
One may want, however, to hold onto the basic tenets of the outward approach without applying a highly revisionist metaphysics. A suggestion in this direction might be along the following lines. There is the social property of being unmarried, and the natural (biological) property of being a man. There is also a property of our language(s), that is, the property of deploying a single term to indicate both properties. English, among other languages, possesses this property. Accordingly, ‘Bachelors are unmarried men’ is contingently true by virtue of the fact that unmarried men are called *bachelors*.

This metalinguistic account of the intuitive privilege of ‘Bachelors are unmarried men’ is slightly different from Fodor’s. Here, the idea is that the property of being a man, and the property of being unmarried, plus a contingent property of our language, would suffice to secure a truth ground for the content in question. Necessity is dropped, but I will claim below that the role of modalities in the explanation of semantic data is a minor one. The idea in common with Fodor’s own version is that intuitively privileged contents are taken to express relations among properties, which we are very familiar with, or attached to, for different reasons. But according to the metalinguistic account, the ontological assumptions that the outward approach needs boil down to two:

- there are symbolic systems, i.e., languages, calendars, algebras, geometries, and the like; these have objective properties, that is, properties that hold independently of our, or anyone’s, opinion on the matter;
- there are societies, ways of life, cultural practices, social conventions, and institutions, and these have objective properties that sciences such as sociology, history, and the like can focus on, and that we ordinarily track and represent with our concepts.

Now, a prima facie objection to the metalinguistic account is that, in general, it does not allow us to describe the alleged special-status content as being metaphysically privileged, that is, expressing necessary relations between properties. In the case we just examined, there is no necessity in the fact that our language calls an unmarried adult man a *bachelor*. Language has a non-negotiable conventional trait, which escapes from natural necessity. So, given the ontological package of linguistic properties it makes use of, the metalinguistic version of the outward approach would deny that special status contents are metaphysically privileged on a general basis.

I acknowledge the point, though I am not quite sure that it qualifies as an problem for the metalinguistic proposal. A supporter of such an approach would still be able to explain that some of the alleged special-status contents are also metaphysically privileged. That would happen in all cases, where linguistic properties were not involved. For example, it is still possible, according to the metalinguistic version, to claim that ‘Nothing is red and green all over’ is metaphysically necessary, because it expresses the metaphysical impossibility of the psychophysical property red’s occupying the same surface as the psychophysical property green, at the same time. Likewise, ‘Schnauzers are dogs’ is about the necessary relation holding between the property dog and the property schnauzer: as a matter of biological impossibility, something cannot be a schnauzer without being a dog.

If some of the alleged special-status contents express necessities while others do not, the class of special-status contents is not homogeneous. This should hardly come as a
surprise, given how the class is individuated – that is, by generalizing on people’s judgmental behavior and habits. But more on this later.

9. Explaining away the illusion

Let us turn now to the second challenge for the outward approach, that is, explaining why special-status contents strike us as special, even though all contents are on a par insofar as they are factual. This is a challenge that the outward approach must face, both in the rich Fodorian version, and in the more sober metalinguistic version sketched above.

A number of solutions have already been proposed. First, there is the idea of theoretical centrality. The idea, raised by Quine (1951), is that contents that seem to be true come what may are in fact very unlikely to be revised in the light of new experience: they are so central to our theory of the world that to abandon them would have a dramatic impact on many other beliefs we hold. For example, contents expressing the relations between concepts such as force, mass, and other basic concepts of physics are extremely unlikely to be questioned. Basic categorical judgments might join this group.

Theoretical centrality is not a cognitive privilege of some contents in the sense outlined above. Theoretically central contents are not so by virtue of the structure of our conceptual system, nor are they true solely by virtue of that; like all other contents, they are true partially by virtue of what their component terms mean, and partially by virtue of what is the case. But they also play a special role in our theory, or theories, for pragmatic reasons, and this suffices to explain why we intuitively distinguish them from other contents. They are, we might say, merely pragmatically privileged.

As many have argued, however, the appeal to theoretical centrality is plausible only for a limited class of cases, which do not exhaust the whole class of semantic data. For example, no one could convincingly argue that ‘Bachelors are unmarried men’ is usually taken as true come what may because it is theoretically central. In fact, Quine’s examples of theoretical centrality are the principles of logic and the laws of physics. The main objection to the appeal to theoretical centrality is therefore that it cannot be a general, all-purpose explanation of semantic data. If there is such a unique explanation of the phenomenon, centrality cannot be it.

A second proposal comes from Putnam (1983), and has recently been discussed again by Fodor (1998, pp. 82–84). In order to accommodate intuitions about the ‘Bachelors are unmarried’ kind of contents, Fodor employs Putnam’s idea of one-criterionhood. The starting point is the claim that some concepts have just one criterion of application, which we happen to share. For example, as a matter of fact, in order to find out whether [BACHELOR] applies to an individual, we always, or at least usually, have to check whether [UNMARRIED] and [MAN] apply to him. This, according to Fodor, would give the impression that [UNMARRIED] and [MAN] are parts of the concept [BACHELOR], and that the content ‘Bachelors are unmarried men’ tells us something about the structure of concepts. And it would suffice to explain the special status we tend to grant to such content. But according to externalist theories, criteria for application are not parts
of concepts; concepts are individuated by what they actually refer to, not by what we happen to apply them to. Therefore, from the fact that we apply [BACHELOR] after checking whether [UNMARRIED] and [MAN] apply, it does not follow that the concept [BACHELOR] itself contains [UNMARRIED] and [MAN]. Semantic data would therefore be explained without resorting to cognitive privileges, but only to contingent facts about our categorization practices.

Against this strategy, it has been objected that the very premise of one-criterionhood is disputable, as strictly speaking no concepts have a unique and widely shared criterion of application. Take the bachelor case again. It can be argued that there are many criteria for checking whether [BACHELOR] applies to someone, from checking the presence of a ring on his ring finger, to monitoring his social life and weekend agenda [Laurence and Margolis (2003)].

The second difficulty with the appeal to one-criterionhood is that, in order to establish that a concept has one criterion, one needs a way to count criteria. It seems, however, that the only way to count criteria would involve the notion of synonymy – for example, being not married and being unmarried would count as one because they are synonyms. But if not married and unmarried are synonyms, then ‘If Ann is unmarried, then Ann is not married’ is true by virtue of the meaning of the terms involved, which was precisely what the outward approach was trying to explain away.

Why not simply say, then, that special-status contents are simply contents that we tend to believe tenaciously and with a high degree of certainty? Such a general characterization would describe the bachelor case, as well as the mass-force case and the color case. Here, the outward approach to semantic data would consist in saying that though the target contents are about properties – colors, dogs, people – they seem to be about concepts because they are so obvious, and we hold to them so tenaciously that anyone who understands the relevant concepts is likely to know them to be true. They express some of the basic facts we grow up with.

The problem with this proposal concerning the outward approach is parallel to the objection raised about centrality. As Laurence and Margolis (2003) have argued, obviousness and tenacity do not guarantee seeming analyticity and seeming apriority. For example, it is obvious that the sky is blue and that stones are hard, but no one claims that these contents have a special status in the traditional philosophical sense. So the class of the obvious and the tenaciously believed is not identical to the class of semantic data, because it is larger. Again, if it is a unique explanation of semantic data that we are after, tenacity of belief and obviousness will not do.

10. A “mixed bag”

It might seem that if the above-mentioned objections are valid, the outward approach fails to explain the semantic data. On the contrary, I think that the outward approach remains a viable option even though each of the objections contains some truth. Let us see why.
The problem with theoretical centrality is that it cannot accommodate all the alleged cases of special-status contents – for example, the bachelor case. Then, there was the remark that the epistemic attitude of certainty toward such contents does not characterize them uniquely, because there are many other contents that we tenaciously hold to, which are undoubtedly a posteriori and factual. Now, these remarks are true. But they do not amount to objections to the outward approach unless a requirement is added that there must be a characteristic property \( p \) that all special-status contents possess, and that any adequate explanation of semantic data must identify.

A supporter of the outward approach need not accept this requirement, at least not without further argument. She may simply refuse to accept that being special-status is the same as being \( p \), where \( p \) is some unique pragmatic, epistemic, or semantic property of contents. In other words, she may opt to be an eliminativist rather than a reductionist about special-status contents.

Along these lines, the outward approach to semantic data would consist in admitting that different contents may appear to have a special status for different reasons, none of which is necessary and sufficient for bring a special-status content. Thus, some contents may strike us as true come what may because they are theoretically central, others for the opposite reason – that nothing depends on them – and so on. In a nutshell, the outward approach to semantic data is viable when combined with the idea that semantic data are, in Fodor’s words\(^6\), a “mixed bag.” Therefore, converging judgments of apriority and nonfactuality are better explained away on a case-by-case basis; it is no surprise that an explanation of the bachelor case will not apply in the mass-force case.

Consider the following scenario: for someone, who denies that UFOs exist, there is no need to prove that all alleged UFO sightings are in fact due to a single kind of phenomenon \( q \). All she needs to prove is that each episode was caused by some kind of illusion, not that all episodes originate from the same illusion (sometimes it might be asteroids, sometimes huge birds, and sometimes Hollywood).

As for the specific objections to one-criterionhood, I think that they can be resisted as well. Therefore, one-criterion concepts may well be one of the kinds one would find in the mixed bag of semantic data. It is undoubtedly true, as the objection goes, that people may employ different sorts of criteria in order to ascertain whether someone is a bachelor. But in fact there is only one conclusive method of doing so: finding out whether the person in question is (male and adult and) not married. Someone who satisfies this criterion is automatically a bachelor; the other criteria (social life, etc.) are at best clues to his marital status.

This is tantamount to saying that bachelor is defined by its unique criterion. Is this something that a supporter of the outward approach could accept? The same worry is contained in the second objection to one-criterionhood raised above. The objection was that, in order to ascribe one-criterionhood, you need to count criteria, but you cannot

---

\(^6\) Fodor introduces the idea (1998, Ch. 4) but then he employs one-criterionhood (as he calls it) as the main explanation.
count criteria without appealing to synonymy – and is synonymy not too much of a concession to the tradition of conceptual analysis?

I think there is no reason why a naturalist philosopher who adopts the outward approach to semantic data should not allow for definitions and synonymy, if these notions are considered in a restricted sense. The idea here is that one can be a factualist about definitions, and say that the few definitions we can point to (including ‘Bachelors are unmarried adult men’) are true by virtue of the properties of our system of representation – in this case, by virtue of the fact that English has one word, bachelor, to indicate the properties of being adult, unmarried, and male. Insofar as languages are nothing but objects of some sort, partially artifacts and partially innate, whose properties can be discovered and theorized about via the scientific method, synonymy and definitions may cease to be forbidden to the naturalist. Moreover, to say that (some) words can be defined – for example, some scientific terms and abbreviations – is not yet to endorse a definitional theory of concepts. Purely linguistic analyticity, as one might call it, is not in itself a commitment to cognitive privileges in the sense of traditional conceptual analysis.

11. Conclusion

The discussion in the first half of this chapter showed that the agenda of the inward approach contains at least the following two points. First, more empirical evidence needs to be found for the hypothesis of hard-wired conceptual rules, which would supplement the transcendental arguments given so far. Second, and most important, a supporter of the inward approach appears to be forced to choose between two alternative strategies. The first is to admit that conceptual relations are merely the rules of a system of representation; the second is to strive for a new version of the Transparency Thesis.

On the other hand, I argued that the main challenge left open to the outward approach is the justification of an ontology, which must be flexible enough to provide truth grounds for all factual contents. And there are reasons to think that this challenge could be met. I also said that the second challenge for the outward approach – that of explaining away the illusion of special-status contents – can be scaled down to a feasible task by dropping the illusory constraint that a uniform solution had to be found for all cases. Nothing so far seems to suggest that the inward approach is better than the outward approach as an explanation of the so-called “semantic data.”

References

This page intentionally left blank
A few years ago, a well-known philosopher of cognitive science, Tim van Gelder, had an existential crisis of sorts. The crisis was brought about by a cognitive scientist who told van Gelder that philosophers had been left off a conference program because the organizing committee felt they did not have anything to contribute to the discipline. What was it, in fact, van Gelder painfully asked himself, that philosophers, given their training and the methods they master, can contribute to cognitive science?

To answer his own question, and resolve his existential crisis, he came up with a number of roles that philosophers were well positioned to play within cognitive science. The papers in this section were all written by philosophers and, taken together, they illustrate the depth and extent of the possible interactions between philosophy and the cognitive sciences.

One of the roles identified by van Gelder is that of the Building Inspector. As the name suggests, the philosopher–inspector works to ensure that the theories and explanations constructed by cognitive scientists respect the basic building codes of proper theory construction. Claude Panaccio’s chapter focuses on one such norm derived from a traditional philosophical doctrine, indeed a doctrine that rose to prominence in the Middle Ages: nominalism. His point is not that cognitive scientists must submit to nominalist principles but that a minimal sensitivity to such principles, in particular the principle called Ockham’s Cleaver, will help prevent needless confusion.

In philosophy, ontology is the field concerned with “what there is” – with the basic furniture of the world, as it is sometimes put – and nominalism is the claim that only singular objects, or individuals, exist: the cup in front of me, those in my cupboard, those in yours, and so on for all the cups in the world, individually. The same is true of all other singular objects: individual atoms, individual animals, individual cars, individual planets, stars, galaxies, and so on. When the thesis is stated in this way, cognitive scientists, at least those who have not been too infected by philosophical theorizing, will no doubt say: “Well, of course, there are only individual things in the world.” However, as soon as they start devising their theories, cognitive scientists get in trouble with nominalism and betray their basic endorsement of the doctrine. What is the concept [CUP]? One cognitive scientist might say that [CUP] constitutes the necessary and sufficient conditions that pick up “cup-hood” or “cup-ness,” that property shared by all cups. But according to nominalism, “cup-hood” or “cup-ness” is not a thing that can be grasped by minds or shared by cups, for it does not exist in the world. Remember that ontology is about what there is and nominalism claims that “cup-hood” and “cup-ness” are not there to be grasped or shared. Mindful of the nominalist position on properties, another cognitive scientist, this one inspired by the brain metaphor, might say that [CUP] is that attractor (in the neural network activation space) that picks up the correlation of visual features exemplified by cups. But again, there are no such things in the world as correlations of visual features. There are individual objects with their visual features, but no correlations of visual features to activate an attractor in a hidden-unit space.

Panaccio himself is a nominalist but his aim is not to sell nominalism (although his chapter contains a short section defending the option on epistemological grounds). Rather, he does two things in his chapter that concern us. First, he shows how a “nominalist
sensibility” can be heuristic for cognitive scientists working on concepts. He therefore introduces what, following Boler (1985), he calls *Ockham’s Cleaver*, a principle that states that one should be careful not to confuse the properties of signifier – in our case, mental representations of concepts – with the properties of what they signify or are about – in our case, categories of individuals. He points to a number of confusions in the literature created by a lack of sensitivity to the Cleaver principle.

Second, he states and defends two constraints that any nominalist theory of concepts should strive to respect. The first constraint is that *the only possible referents of concepts are individuals*. Some concepts refer to no individuals at all (e.g., [FLYING HORSE]); some refer to one individual (e.g., [THE PRESENT QUEEN OF ENGLAND]); and some to several individuals (e.g., [CAR]). Mindful of this constraint, cognitive scientists should refrain from taking the referents of the above-mentioned concepts as, respectively, the empty set, the singleton containing the present queen of England, and the set of all cars. Concepts should not be taken to refer to properties, attributes, features, or sets. It might be argued that in the case of sets, non-nominalist talk can easily be translated into a form that respects this constraint, but the nominalist would reply that it is nevertheless better to stick to the principle and refrain from talking of sets as referents of concepts: such a regimen may well prevent confusion from creeping in through the back door. Panaccio’s second constraint concerns the mental representations of concepts. They, too, are individuals (e.g., individual brain states). Hence, they are not the sort of things that can be public or shared, as some cognitive scientists would have it.

Another role recognized by van Gelder for the philosopher of cognitive science is that of *Pioneer*. Throughout history, philosopher-pioneers have been both admired and mocked: admired when their work led to some new science, political system, or ethical principle; mocked when, as may happen to all pioneers, they go astray, get lost, or vanish without a trace (think of all those pioneers who lost their lives in the search for the Northwest Passage). Chris Eliasmith is one among a few philosopher-pioneers defending the view that the study of categorization, or semantics generally, must begin with the neurocomputational study of the cognitive brain. Only time, and much work, will tell if a *computational cognitive neuroscience of meaning* is a foolhardy enterprise or if it turns out to be our first real scientific foothold in semantics. Whatever happens, Eliasmith’s work (here and elsewhere) is surely one of the best attempts at finding the Northwest Passage to Naturalized Semantics.

It should be noted that that the expression “computational neuroscience” has two meanings (even computational neuroscientists get them confused). These meanings are not incompatible, to be sure, but one is *stronger and includes* the other. The weaker notion is common in science generally, when one names a new discipline or method *computational so-and-so* (where “so-and-so” nowadays can almost be any scientific discipline imaginable – computational chemistry, computational physics, computational biology, computational meteorology, computational genetics, and so on). In this more common sense, the expression simply refers to the use of mathematical modeling and computer simulation to understand the set of phenomena relevant to that field. Unlike
other scientific disciplines (except engineering), however, neuroscience also affords a stronger meaning, whereby the mathematical modeling and computer simulation referred to in the weaker notion are used to understand the systems studied by neuroscience as computational devices. Unlike the atmosphere, as studied by computational meteorology, the brain as a whole – but also its components (various brains systems), their components (neural assemblies), and perhaps even their components (neurons) – is understood by proponents of this stronger view of computational neuroscience as a hierarchy of nested computational systems. The computational neuroscience of vision, for instance, will understand visual phenomena such as color or motion perception as computational phenomena, and understand the brain systems devoted to vision as computational systems devoted to generating these phenomena. To mark the difference between the two expressions, I shall henceforth capitalize the stronger notion. It is important to emphasize that these two senses of the expression have different epistemic statuses. The weaker notion simply reflects the plain methodological fact that, given the power of today’s computers, almost any physical phenomenon is amenable to computer simulation, even those that are very computationally complex (think of computational fluid dynamics). The epistemic status of the weaker notion is thus quite straightforward: it is an observed sociohistorical fact regarding the growing use of a particular set of tools in today’s science. By contrast, the stronger notion is a bold, and some believe quite false, hypothesis about the nature of cognition according to which the brain is a hierarchy of computer systems. For that hypothesis to be true, some structures in the brain must be representations, that is, they must possess semantic properties. Computing is the controlled transformation of structures that stand for the arguments of a function into structures that stand for its value. If a semantics mapping brain structures onto the arguments and values of a function can be described, then these structures become representations, the neural stand-ins of these arguments and values. If it turns out, however, that no semantics can be given for brain structures, then Computational Neuroscience is false.

In that context, Eliasmith’s chapter plays a double role. The first role is explicitly stated by the author. He argues that, whereas traditional accounts of categorization have all addressed the phenomenon from the point of view of psychology or linguistics, it should now be tackled from the point of view of computational neuroscience (not capitalized). In that view, categorization, like neuronal integration, motion perception, and movement coordination, is a neural phenomenon that can and should be mathematically modeled and simulated on computers. With the demise of the strong reading of functionalism in the philosophy of cognitive science, which entailed the absolute autonomy of the psychological level of explanation with respect to the neurological level (the latter responsible for “mere implementation” issues), no one will object to a computational neuroscience of categorization. There are, however, many ways in which this could be achieved and the one Eliasmith chooses reflects the second role of his chapter. He attempts to provide a computational neuroscience of categorization by providing a semantics for a class of transient neural states (activation of neuronal populations). Such a semantics would show that the neural states are representations and, thus, that Computational Neuroscience is well-founded.
A further role identified by van Gelder for the philosopher of cognitive science is that of Cartographer. The philosopher–cartographer aims to show how the various theories, explanations and hypotheses that make up cognitive science, or indeed science in general, hang together (or not). This is what Edouard Machery and Luc Faucher do. They explore how two sets of theories of categorization, that is, social constructionist and cognitive evolutionary theories of categorization, can be integrated into a coherent framework.

To do this, they address a phenomenon that has largely been neglected by cognitive scientists working on categorization: the fact that some typologies (ways of dividing a domain into categories) seem more natural than others. According to most accounts of categorization, or cognition generally, the phenomenon should derive from the culture or physical world one is brought up in (or some mix of the two). Importantly, the phenomenon should not derive from any important fact about cognition. Accordingly, its study should not reveal anything important about the mind and hence cognitive scientists should ignore the phenomenon. With the rise of the evolutionary attitude in cognitive science, however, the phenomenon takes on a new urgency since it may reveal important fractures in our cognitive system: some typologies seem more natural because they form the representational core on which a cognitive module, i.e., a domain-specific cognitive computational system, is built. By contrast, those typologies that do not seem natural are this way because they are learned by some slow domain-general system or, worse, they must co-opt some module’s own representational core.

Instead of tackling this important issue in the abstract, Machery and Faucher focus on one such typology (or set thereof): those that underlie racialist thinking, that is, the concept [RACE], and various racial concepts such as [BLACK], [WHITE], [HUTU], [TUTSI], and so on. Most scientists today (and certainly all those reviewed by Machery and Faucher) agree that racial concepts have no biological reality, that the current human species (Homo sapiens sapiens) cannot be biologically divided into subcategories. The question, then, is where do such concepts come from? Until recently, racial categories had been ignored by cognitive scientists and were mainly studied by social psychologists, anthropologists, and philosophers, who viewed racial categories as social or cultural constructs that were self-fulfilling, transient, culture-specific, time-specific, and socially or politically motivated. The social constructivists have recently been opposed by a group of cognitive scientists inspired by the theory of evolution who claim that racial concepts are evidence of a domain-specific cognitive module. Evolutionary psychologists agree that there is no race module and that racial concepts are rather a by-product of the functioning of some other module, the nature of which is still in dispute (Machery and Faucher survey the important contenders). The structure of the debate is clear: social constructivism on one side and a number of evolutionary psychological theories on the other. As cartographers, Machery and Faucher weigh the merits and defects of each, ending their paper with a number of requisites for future theories of racialism that integrate the best contributions from each of the traditional adversaries.

Although they open their chapter by stating that they will address the fact that some ways of dividing a domain into categories seem more natural than others, they do not
return to this issue, but stay focused instead on the categories that serve for illustrative purposes. It would have been nice if they had taken the time to state that position properly: Why do some ways of dividing the world seem more natural than others? Do all categories behave the way racial concepts do? Of course, one can read between the lines and infer to a great extent what their probable position is. Social constructionists presuppose that racial concepts are learned by some domain-general cognitive system. Evolutionary psychologists believe that racial concepts are derived through the activity of an innate cognitive module. Neither is completely right or wrong. Racial concepts are created in the interplay between, on the one hand, domain-general mechanisms that incorporate culturally constructed (and even perhaps culturally evolved) factors into the mind and, on the other hand, domain-specific mechanisms that, although they were not selected to build racial concepts, favor and constrain the formation of such concepts. As promising as this story sounds, it still needs a mechanism to generate the proper interactions between the domain-general and domain-specific sources of concepts.

Van Gelder’s crisis made him seek out what philosophers could contribute to cognitive science. His crisis may have been worsened by the fact that, over the past 50 years, philosophers have borrowed heavily from cognitive science: in the philosophy of mind, obviously, but also in the philosophy of language, in epistemology, ethics, aesthetics, and so on. In recent years, some authors have also suggested that philosophers could draw on cognitive science to help revive a traditional method of theirs: conceptual analysis.

In the first half of the twentieth century, it was believed that philosophy was not in the business of making claims about the world, but that it should rather concern itself with the clarification of our claims about the world. Our conceptual system, it was believed, is such that, even with the best of intentions, we are sometimes led to make meaningless claims or ask meaningless questions about the world. Philosophy was thus given the job of seeking out those conceptual errors that allowed meaninglessness to infiltrate thought. One way to clarify thought was to translate the statements that express it into first-order predicate logic. Another important way to achieve this goal was conceptual analysis: analyzing the concepts that were thought to be responsible for such errors. This school, quite influential in the English-speaking world, was called Analytical Philosophy.

Conceptual analysis works by analyzing a concept into its constituent concepts. The statements resulting from such work were said to be analytic, that is, true or false solely on the basis of the meaning of their constituent concepts. And analytic statements were taken to be a priori (true or false independently of experience) and necessary (true or false in all possible worlds that contained them if true or false in the actual world). Analytical Philosophy is no longer with us in the form it had in the first half of the twentieth century: it received its coup de grâce when philosophers came to think of conceptual analysis as ill-conceived, that is, when they came to believe that its intended product, analytical sentences, simply did not exist (not, at least, as they were understood and not, more importantly, as they had to be for philosophers to do the work Analytical Philosophy expected of them): no sentence, Willard Quine showed in a famous paper, can be true or false solely by virtue of the meanings of its constituent concepts.
They say that everything old is new again and, as Elisabetta Lalumera reminds us in her chapter, conceptual analysis is currently undergoing something of a revival. Notwithstanding Quine’s pronouncements about the existence of the kind of analytical sentences that were thought to be necessary if Analytical Philosophy was to be in the thought-clarification business, the fact remains that some statements do intuitively appear to all competent speakers of a language to be analytical and necessary (the classical example being *Bachelors are unmarried*). What is more, people do intuitively confer a certain privileged epistemic status upon such statements. Facts such as these surely deserve an explanation. As Lalumera notes, there are two attitudes that naturalistically minded philosophers, that is, philosophers who want their explanations to be consistent with the best science of the day, may take toward this fact. She calls these two attitudes the *inward* and the *outward approach*. Proponents of the inward approach believe that the intuitions to be explained reflect something about the structure of our cognitive system. If that is true, then an important part of the explanation will be provided by cognitive science. Proponents of the outward approach believe that these facts reflect something about the structure of the world. After raising some problems with the inward approach, Lalumera defends her own brand of the outward approach whereby “relations among concepts, whether innate or acquired, mirror the relations among the real-world properties they refer to” p. (1055). Note that, as stated, Lalumera’s position conflicts with Panaccio’s nominalism, according to which there are no properties in the world, let alone relations between properties, for concepts to mirror. This, perhaps, illustrates one last prominent role of philosophers: disagreeing with each other. Whether or not readers are convinced by the arguments contained in these four papers, I believe that they cannot fail to be impressed by the role philosophers can play in cognitive science.
This page intentionally left blank
SUBJECT INDEX

abstraction 30–31, 36, 78, 158, 645, 704
acquisition
  of concepts 253, 313
  of language 37, 257, 358, 444, 485, 488, 536
  of lexicon 466, 471
  of phonemic categories 510, 604
  of phonetic categories 497, 510, 511, 538
  of second language 498, 515
  of semantic categories 459, 540
  of syntactic categories 359, 515, 519
adjective 57, 59, 169, 330, 430, 437
adverb 57, 169, 383, 527
agent
  adaptive 719, 720, 729, 734, 735
  reactive 743, 747–749
agreement
  atomic/singular 388
  collective/plural 388
AI see artificial intelligence
AL see artificial life
aliasing 744, 745, 746
ambiguity 75, 76, 340, 745, 900
Amerindian languages 54, 61, 483
amnesia
  anterograde 127
analogy 50, 150, 156, 157, 158
analytic
  and nonanalytic 127, 128, 129, 130
anterior cingulate 130, 133, 135–136, 552
animat 745, 746
aphasia 605
apraxia 132, 136, 605, 606
argument 52, 411, 438, 701–705
articulatory system 389
artificial intelligence (AI) 26, 141, 142, 160, 795, 893, 957
artificial life (AL) 26, 741
artificial neuron network 825, 827, 837
association rule 367, 376–377, 921–923
atoms 243, 246
attention 103, 131, 132, 270, 451, 553–555
basal ganglia
  caudate 128, 559
biological motion 134
bootstrapping 260, 436–438, 448, 450
borrowing
  lexical 61, 65
Broca’s area 605–610
carrier 67, 1045
categorical perception 25–26, 34–35, 602, 610, 721–722
category
  auditory 497, 538
  category-specific disorder 122
  formation 448, 497, 869
  functional 327–329, 384, 516–517, 519–524, 758
  grammatical 45, 434–435, 436–438, 449, 482–486
innate 23
learning 23–24, 547, 729
lexical 47, 169, 403–404, 517–519
meaning 52, 252, 471–472, 849, 1036
naming 724–726, 730, 731, 734
nonspeech 502–506
perceptual 25–26, 498, 547, 573, 651, 719
phonetic 497, 537
phonological 61, 383, 610
semantic 168, 172–175, 184, 459
speech 506–510, 601
syntactic 47, 330, 401, 423, 481, 515
universal 94, 396, 472–475
categorization
  classical view 245, 279, 296
  embodied 739
  judgments 288
  infant 481
  perceptual 498, 547, 574, 601, 651, 1037
Subject Index

phonetic 486, 499
probabilistic task 127–128, 281
racial 1010, 1021
semantic 167, 436–438, 540
categorizer
  analogizing 759–761
  emulating 753–755
  linguistic 765
caudate nucleus 551, 556, 560
cerebellum 601, 611–612
child language 348, 358–359
chroma 225
CIE L*A*B* 235
classification
  ambiguity 900
  classifiers 113, 528, 967, 968
  clustering 27, 136, 235, 510
coalitional module 1017, 1021–1025
cognition
  embodied 723, 731, 733, 740, 741
  ethnic 1024–1029
  infant 436
  situated 680, 960
cognitive symbol grounding 720–723
collocations 945–946
configurationality 417
color
  basic terms 107, 225, 228
  categorical perception of 25–26
  classification 105–109
  naming 108, 234–240
  space 106, 235, 839
  terminology system 225
Common KADS 980
competence 49, 53, 77, 83, 159, 385, 706–709, 710–711
completeness 980, 984
complex systems 345, 989
computational
  learning theory 987
  modeling 565, 652, 674, 675, 735
  neuroscience 761–762, 1037
computer 141, 955
concept
  abstract 645–646
  acquisition of 471, 680, 682
  classical view of 73–75, 279
  cultural niches of 1013, 1029
  empty 71
  existential vs. purely intentional usage 75–76
  development of 244, 249–252, 482
  inductive recognition of 935
  internal role of 82–87
  learning 126, 652, 653–664, 675, 823–843
  linguistic instance of 937, 938, 947, 952
  moral 683, 693
  one-criterion 86, 1067, 1068, 1069
  perception 149–150, 620, 653–664, 680, 825
  prototype view of 997
  recognition in texts 936, 944–951
  relational 937, 1003–1004
costual
  analysis 969, 1055
  change 243, 257
  formal 919, 920–921
  graph formalism 982
  lattice 861–862
  primitives 146, 257, 271
  situated 619
  space 152, 827–829, 830–831, 832–834, 834–836, 837
  substance 389, 391
  system 189, 620, 621–629, 629, 638
  conjunctive task 665, 668, 670, 671
  conjunction 64, 169, 367
  connectionism 142, 680, 726, 825, 827
  connectionist models 450, 663, 724–729, 733, 734
  connector 175, 176, 181
  connotation 1003, 1004
  consensus level 231, 232
  constituent 65, 136, 173, 369, 384, 1039, 1078
  constituency
    rule 177, 179, 180, 181
    tree 177, 180, 181
  constraints of linearization 389
  constructionism 1010, 1011, 1014, 1015
  convexity 831, 833
  coordination 177–179, 611, 830
  coordinator 175–181
  copula 59, 430, 857
  corrective inference 713–715
  correlated dimensions 664, 665, 667
  correlational structure 100, 115
  cultural
    pattern 93, 105, 115
    transmission 1025, 1028
    variation 109, 115, 199, 1028
    cumulativity 170, 171, 183
  data mining 863, 911, 917–918, 927
coding 1040–1041
denotation
  multiple 1002
determiner
  definite 387–388, 397
deployment 53, 154, 244, 249–252, 257, 522–524, 535
differentiation 663, 673
dimension
  quality 828, 829–831
  relevant 504, 505, 509
  irrelevant 588, 590–591, 596, 664
discourse orientation 414, 418
discrimination 32–33, 461, 488, 498, 603, 876, 880
disease
  Huntington’s 127, 575
  Parkinson’s 127, 555, 563, 575
  schizophrenia 132, 136, 563
disjunctive problem 1050–1052
dispersion 235–237
divisivity 170–172
duration 131, 132, 502
dynamic systems 22, 628, 643, 715–716, 746
dysarthria 611
effects
  top-down 134, 144, 146, 148, 722, 786
  embedding 78, 405
  embodiment 635–636, 729, 734, 742, 870
  empiricism 679
  endocentricity 50
  ERP see event-related potential
error 83, 290, 709–710, 710–711
  essentialism
    psychological 283, 1020, 1027
  ethnomethodology 109
  event-related potential (ERP) 121, 122, 134, 491
  externalism
    strong vs. weak 78
family resemblance 40, 41, 73, 129, 306
  feature
    adaptation 28, 668–669, 671, 716
    creation 674
    diagnosticity of 804–805
    distributed network of 123
    hypothesis 142, 523
    phi-f. 10, 411, 429
    psychological 396, 672
    selection 32, 33–34
    strength 516, 520, 524
    weighting 32
feedback 24, 26
  filter task 664, 665, 667, 668
  fMRI see functional magnetic resonance imaging
  focal points 981
  focus task 225, 230
  folk
    biology 689, 758, 1025–1026, 1028
    sociology 1017, 1019, 1021
    formant frequency 502, 506, 507
    Fregean semantics 1005, 1039
    functional magnetic resonance imaging (fMRI) 121, 128, 134, 632, 774
    fusiform face area (FFA) 123, 125
  fuzzy set 147, 893, 896, 897, 899, 902
Galois lattices 846, 859, 862, 920
gapping 176, 324, Garner
tasks 664
  interference 664–665
gender 57, 80, 87, 393–394, 396, 483, 528–529
graph structures 981, 982, 988
  first-order
  grounding
    sensorimotor 36–37, 721
    symbol 36–37, 719
headingness 368–369, 370, 371
hippocampus 128, 555, 581, 751
imagery 160, 622, 623, 632, 686, 758
induction 153, 155–157, 442, 703–705, 709, 947–948, 936, 951
inference 154, 627, 628, 638, 639–643, 701, 752, 759, 838
inferior temporal cortex 120
inflection 96, 355, 441–444, 520, 524
information
  integration 560–563, 566–567, 576
  theory 35, 1040
innateness 688–691, 1062
insula 605–607, 612–613
intentionality 78, 759
interjection 63–65, 169
interlanguage grammar 515, 516, 517, 519, 522, 523, 528
invariance 21, 23, 26, 28, 31–32
isomorphism 982, 1039
itemset 921–923, 925, 926–927, 928
Subject Index

judgment 72, 74, 84
modal 72, 74, 84
part/whole 184
similarity 103, 112, 126, 127, 195, 277

kinship terminology 101–105
knowledge
approach 52, 243, 245, 251,
k.-based technology 979, 980, 989
discovery 846, 913
elicitation 988
modeling 979

L2 see second language
language
acquisition 257, 358, 444, 485, 488–492, 726
artificial 439, 441, 490, 732
associational 347, 424
contact 61–67
diversity 45
evolution 729–733, 735
formal representation of 981
grounding 719, 723, 724, 729, 731
isolating 350–351, 360, 486
monocategorial 350–351, 358, 365, 424
natural 23, 259, 260, 344, 355, 360, 831, 1038
of thought 682, 995, 1038–1039
spoken 401, 425, 428
typology 359
see also sign language; second language
law
noncontradiction 895, 896–897, 901, 905, 907
excluded middle 895, 896–97, 901, 907
learning
algorithm 26, 28, 797, 988
cue-based 450
distributional 439, 449–450
explicit 36
general model 26
Hebbian 799–802, 808, 812
instrumental 24–25
heuristic knowledge 984–985
implicit 559–560, 565–567
machine 26, 918, 987
operant 24–25
procedural 548, 556–559, 565, 578
supervised 24, 27–28, 34, 743, 797, 918
unsupervised 26, 27, 34, 819

lexicon 47, 48, 56–58, 61, 66, 337, 466
linguistics
cognitivist 444
generative 168, 321
notionalist 168, 169
structuralist 168, 173
locus
assignment 393, 396
reusing 386, 393
point locus 386, 387–388
zone locus 388
logic
classical 706, 708, 710
magnetoencephalography (MEG) 121, 780
mapping task 257
meaning
theory of 73, 172, 1039
covariance theory of 864, 1039
medial temporal lobe (MTL) 120, 128, 555, 581, 587, 596
MEG see magnetoencephalography
membership
graded 900, 997
memory
auto-associative 794, 797, 800, 813
explicit 581, 582, 587
implicit 482, 579, 930
working 547, 548, 551, 554, 555, 561
mental representation 77, 244, 246, 252, 257, 279, 280
meta-level graph 988
metalinguistic 1066, 1067
metaphor 156, 686, 704, 759, 1044
metonymy 352, 704
mind
philosophy of 691, 1077
theory of 758
mismatch 48–52, 54–55, 61, 67
modality
oral 381, 382, 388
gestural 381, 388, 392
m.-specific systems 621, 622–623, 629, 630–631, 636, 637
modeling
complexity 986–988, 989
in layers 724, 727, 988
modularity 48–52, 735, 959
morphology 66, 168, 228, 332, 339
morphology-before-syntax 519, 522–523
motor
Subject Index

naming
array 229–230

network
collectionist 26, 450, 721, 723, 726, 734, 744–745, 797
neural 26, 131, 142, 245, 450, 498, 536, 720
recurrent 131, 726, 794, 795, 802

neural
computation 1041
dynamics 750
networks 131, 142, 245, 450, 536, 563, 720–724
representation 245, 757, 1039, 1043, 1047–1048, 1051
transformations 1041

neuroimaging
fMRI 121, 125, 128–129, 134, 559
PET 121–122, 128, 759
ERP 121–122, 134, 491

neuron doctrine 120
neurons 107, 120, 622–624, 644–645, 751, 755, 759, 784–785

neuropsychology 768
neuroscience 119, 120–121, 130, 548, 1037
neurosemantics 1039, 1052
nominalism 993–996, 1001, 1002, 1004, 1006, 1013, 1073–1074

nonanalytic 127–130
noun
abstract 172, 269
collective 171–172
common 170, 183, 537
concrete 539, 540
count 170–172, 182–183, 258
grammatical 256, 351, 373, 403, 405, 434–439
mass 167, 170–172, 182–183, 258
number 443, 467
numerals 183, 264
plural 183–184
proper 58, 67
singular 170, 182

Ockham’s Cleaver 995–999
Ockham’s Razor 995

optionality 530

parsing 133, 134, 136, 450–452, 489–490, 659, 662

partonomy 131

pattern
completion 620, 622, 624, 627–628
matchers
perception
categorical 25–26, 538, 602, 604–605, 610–612, 721, 725
of concepts 825
limited 900–901, 904, 907
perceptual
categorization 130, 488, 498, 581, 1016, 1037
organization 662, 671, 675, 710–711
PET see positron emission tomography
philosophical naturalism 1056
phonemes 80, 96, 445, 510, 601–607, 609–611, 613
phonetics
acoustic 78, 481, 482, 485, 492, 536
articulatory? 389, 604, 607, 609
phrase marker 177–181
see also constituency tree
phrase structure rule 177
see also constituency rule
polysemic 632, 945
positron emission tomography (PET) 121, 122, 128
possibility
distribution 905
theory 143
prefrontal cortex 129–130, 133–134, 551, 595, 601, 606, 612
preposition 56–57, 61, 65, 169, 320, 385, 462–463, 483
procedural
knowledge 121, 132, 560
learning 547, 548, 551
processing
analytic 127
explicit 127
implicit 126
infant language 374, 483
nonanalytic 126
projection 39, 58, 146, 158, 320
pronoun
interrogative 182
personal 182
relative 182
property
gradual 896, 904
necessary 279–280, 286
sufficient 279, 287
prototype
classical view 243, 245
theory 77, 147
psychology evolutionary 1016
psychophysical space 497, 498, 538

racialism 1909
radiation
  UV-B 238
real kinds 855
reasoning
  hypothetical 708–709
  nonmonotonic 157, 838–839
recurrence 31–32
recursion 51, 405, 624, 941–942
reduplication 350, 361
reference 148, 170, 171
referent 76–80, 96, 106, 198, 269, 386
region of interest (ROI) 125, 606
reinforcement 34, 36
relativity
  linguistic 761
representation
  existential vs. purely intentional usage 75–76
  formal 394
  levels of 838
  mental 77, 244–246, 252, 279
  relation 75
  space 251
representational
  hierarchy 47
  vehicles 537
robot 154–155, 159–160, 690, 729, 731, 732–735
robotics 21, 721, 731–732, 741, 749
rule
  contrast sensitivity 179
  interpretation 179
  r.-based 154, 549

schizophrenia 132, 136, 563
second language (L2) 515
  acquisition 519–521
segmentation 490, 491, 492
  figure-ground 785
selection 32, 33, 284
self-organization 869
semantics 167
semiotics 47, 51, 354–356
sense 29, 36, 40, 133, 173
sensitivity 124
sensitization 663
sensorimotor 21
  abilities 21
  coordination 746–747

shared attributes 998
sign language
  Quebec Sign Language (LSQ) 381
similarity 297–300, 386–388, 897
simulation 635, 757, 758–759
singular 1002–1006
  see also number
skill 24, 29, 127, 136, 475, 818
space 461–464
  psychophysical 497, 498, 500, 502, 511
  search 450–451
speech
  parts of 169, 350, 434, 449
  perception 601
  production 611–612
statistical dependence hypothesis 1036, 1046, 1047, 1048, 1051, 1052
structured event complex (SEC) 134
subcategorization 326, 948
subcategory 132, 336, 447, 448, 711, 839
subordination 181, 405
subordinator 175–181
substance
  perceptual 389, 395
  conceptual 389
supervision 24, 26, 27, 36, 497, 499, 501, 506
switching 553–555
syllogism 144, 145, 706, 707, 708, 710
symbol
  grounding 719
symbolic
  data analysis 846
  representations 47, 825, 826, 828, 836, 838, 840
syntax 47, 50, 51, 53, 66
system
  dynamic 715

template 40, 525, 988
temporal–parietal junction 609
temporal response density (TRD) 134, 135
tense 327, 363, 516
text
  analysis 955
  mining 935
  standardization 940
tagging 940
terminology of 944
thalamus 552, 556, 611
theory
  category 40
  generative grammar 321, 325, 350
  grammatical 52
<table>
<thead>
<tr>
<th>Subject Index</th>
<th>1087</th>
</tr>
</thead>
<tbody>
<tr>
<td>of language 347–377</td>
<td></td>
</tr>
<tr>
<td>of mind</td>
<td></td>
</tr>
<tr>
<td>“theory” t 278, 296, 297</td>
<td></td>
</tr>
<tr>
<td>token 42, 603, 608</td>
<td></td>
</tr>
<tr>
<td>type</td>
<td></td>
</tr>
<tr>
<td>typicality 40, 41, 115, 131, 643, 895, 898, 899, 906, 997, 1060</td>
<td></td>
</tr>
<tr>
<td>uncertainty 35</td>
<td></td>
</tr>
<tr>
<td>universals</td>
<td></td>
</tr>
<tr>
<td>category 396–397, 473</td>
<td></td>
</tr>
<tr>
<td>color 234–240</td>
<td></td>
</tr>
<tr>
<td>emotion terms 187</td>
<td></td>
</tr>
<tr>
<td>language 45</td>
<td></td>
</tr>
<tr>
<td>semantic 167</td>
<td></td>
</tr>
<tr>
<td>vagueness 891</td>
<td></td>
</tr>
<tr>
<td>value 567, 587, 611, 665, 798</td>
<td></td>
</tr>
<tr>
<td>variability 234, 302, 414, 583, 585, 590, 592, 1012</td>
<td></td>
</tr>
<tr>
<td>verb</td>
<td></td>
</tr>
<tr>
<td>agreement 408</td>
<td></td>
</tr>
<tr>
<td>raising 521, 525, 526, 529</td>
<td></td>
</tr>
<tr>
<td>Voronoi tessellation 833, 834, 835, 836, 841</td>
<td></td>
</tr>
<tr>
<td>WCS see World Color Survey</td>
<td></td>
</tr>
<tr>
<td>Whorfian hypothesis 34, 35</td>
<td></td>
</tr>
<tr>
<td>World Color Survey (WCS) 223</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td></td>
</tr>
<tr>
<td>class 447–449</td>
<td></td>
</tr>
<tr>
<td>order 413–417</td>
<td></td>
</tr>
</tbody>
</table>
Plate 1.3  (a). An artificial-life simulation of mushroom foragers. Mushroom-categories could be learned in two different ways, by sensorimotor “toil” (trial-and-error learning with feedback from the consequences of errors) or linguistic “theft” (learning from overhearing the category described; hearsay). (b). Within a very few generations the linguistics “thieves” out-survive and out-reproduce the sensorimotor toilers. (But note that the linguistically based categories must be grounded in sensorimotor categories: it cannot be theft all the way down.) [Reproduced with permission from Cangelosi and Harnad (2001)].
Plate 5.5(a). Main effects of BOLD showing significant activation in caudate, anterior cingulate, medial frontal gyrus and anterior cingulate over 40 subjects during three blocks of categorization decisions in the condensation task. (b). Interaction effect between priming and task over 40 subjects during three blocks of categorization decision in both condensation and filtration task. Significant BOLD activation is shown in both posterior cingulate and parahippocampal gyrus.

Plate 5.7 Contrast between event change points determined by significant temporal response density (TRD) points and nonevent points. Typical areas that appear engaged are the anterior cingulate, the inferior frontal gyri, the middle frontal gyri, the precuneus, the parietal lobule, and the dorsal lateral prefrontal cortex. These are voxel activities that are used for further clustering and graph analysis.
Plate 6.5. Evolution of cognition.

Plate 9.1  (a) The WCS stimulus array. (b) Munsell and WCS coordinates for stimulus palette. The leftmost column and the top row give the WCS coordinates for lightness and hue, respectively. The rightmost column and the bottom two rows give the Munsell coordinates for Value and Hue, respectively. Entries in the body of the table show the corresponding Munsell Chroma numbers. [With regard to the A and J rows, there are no Munsell hues at the extremes of Value (lightness): 9.5 (white) and 1.5 (black).]
Plate 9.4  (a) Contour plot of WCS speakers’ naming centroids, compared with English naming centroids (black dots). (Source for English naming centroids: Sturges and Whitfield 1995.) The outermost contour represents a height of 100 centroids, and each subsequent contour represents an increment in height of 100 centroids. Source: Kay and Regier (2003). (b) Monte Carlo test for clustering within the WCS data. The distribution of dispersion values shown in blue was obtained from 1000 randomized datasets. The red arrow indicates the dispersion value obtained from the WCS data. [Source: Kay and Regier (2003)].
Plate 9.5. Contour plot of WCS chromatic focus peaks compared with English naming centroids. [Source for English naming centroids: Sturges and Whitfield (1995)].

Plate 9.6. Contour plot showing the distribution, over chromatic stimuli, of best examples of grue terms in the WCS. Outermost contour represents a height of 10 hits; each subsequent inner contour represents a height increment of 10 hits. [Source: Regier and Kay (2004)].
Plate 27.1. (a) Representation of how the onset frequencies of the second and third formant can be manipulated to create a continuum of syllables that range from “ba” to “da”. (b) Idealized representation of how listeners perceive items along this syllable continuum.
Plate 34.1. (a) The four scenes used in this experiment and the corresponding low-level category names ("field," "mountain," "desert," and "dune") that were learned by all participants, and the high-level categorizations ("flat" and "hilly") that were learned by participants given luminance cues (LUMI participants). The two histograms illustrate the RT curves predicted by a SLIP categorizer model. (b) The proportion of the four error types as a function of presentation time for each observer group. The solid and dashed lines are, respectively, the average best-fit curves from individual data points applying the SLIP model to the LUMI participants and to participants given chrominance cues (CHRO participants). The black curves represent the proportion of errors on none of the perceptual dimensions (e.g., respond "field" when presented with a field scene); the green curves represent the proportion of errors on the luminance dimension (e.g., respond "mountain" when presented with a field scene); the red curves represent the proportion of errors on the chrominance dimension (e.g., respond "desert" when presented with a field scene); and the blue curves represent the proportion of errors on all dimensions (e.g., respond "dune" when presented with a field).

Plate 35.6. Eigenvalue spectrum resulting from using the Eidos model in the letter-classification task. Eigenvalues are displayed in decreasing order.
Plate 35.11. Eigenvalue spectrum developed using the learning/unlearning procedure with the Eidos model. Eigenvalues are displayed in decreasing order.

Plate 35.12. Eigenvalue spectrum developed after 10,000 learning trials in the letter-classification task. The eigenvalues are displayed in decreasing order and the frequency is expressed as the ratio of the number of unlearning trials to the number of learning trials.
Plate 35.13. Eigenvalue spectrum developed after 10,000 learning trials in the letter-classification task. The eigenvalues are displayed in decreasing order and the pattern of unlearning represents the type of sequences used to train the network. For example, 10/10 signifies a series of 10 learning trials followed by a series of 10 unlearning trials.