Theory and Applications of Natural Language Processing
Edited volumes

Aline Villavicencio
Thierry Poibeau
Anna Korhonen
Afra Alishahi *Editors*

Cognitive Aspects of Computational Language Acquisition

Springer

Theory and Applications
of Natural Language Processing

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

For other titles published in this series, go to
www.springer.com/series/8899

Aline Villavicencio • Thierry Poibeau
Anna Korhonen • Afra Alishahi
Editors

Cognitive Aspects of Computational Language Acquisition

Springer

Editors
Aline Villavicencio
Institute of Informatics
Federal University of Rio Grande do Sul
Porto Alegre, Brazil

Anna Korhonen
Computer Laboratory
University of Cambridge
Cambridge, UK

Thierry Poibeau
Ecole Normale Supérieure
Université Sorbonne Nouvelle
LATTICE-CNRS, Paris, France

Afra Alishahi
Tilburg center for Cognition
and Communication (TiCC)
Tilburg University
Tilburg, The Netherlands

ISSN 2192-032X
ISBN 978-3-642-31862-7
DOI 10.1007/978-3-642-31863-4
Springer Heidelberg New York Dordrecht London

ISSN 2192-0338 (electronic)
ISBN 978-3-642-31863-4 (eBook)

Library of Congress Control Number: 2012954240

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

We would like to acknowledge the support of the labex (cluster of excellence) TransferS (France), the Royal Society and EPSRC grant EP/F030061/1 (UK), CAPES-COFECUB grant 707/11 and CNPq grants 479824/2009-6, 551964/2011-1 478222/2011-4 and 309569/2009-5 (Brazil).

Contents

Computational Modeling as a Methodology for Studying Human Language Learning 1
Thierry Poibeau, Aline Villavicencio, Anna Korhonen, and Afra Alishahi
1 Overview 1
1.1 Theoretical Accounts of Language Modularity and Learnability 2
1.2 Investigations of Linguistic Hypotheses 5
2 Computational Models of Language Learning 7
2.1 What to Expect from a Model 8
2.2 Modeling Frameworks 10
2.3 Research Methods 13
3 Impact of Computational Modeling on the Study of Language 16
4 This Collection 17
4.1 Methods and Tools for Investigating Phonetics and Phonology 17
4.2 Classifying Words and Mapping Them to Meanings 18
4.3 Learning Morphology and Syntax 19
4.4 Linking Syntax to Semantics 20
5 Concluding Remarks 22
References 22

Part I Methods and Tools for Investigating Phonetics and Phonology

Phon: A Computational Basis for Phonological Database Building and Model Testing 29
Yvan Rose, Gregory J. Hedlund, Rod Byrne, Todd Wareham, and Brian MacWhinney
1 Introduction 29
2 The PhonBank Project 31
2.1 PhonBank 31
2.2 Phon 32

3 Phon 32
3.1 Project Management 33
3.2 Media Linkage and Segmentation 34
3.3 Data Transcription 34
3.4 Multiple-Blind Transcription and Transcript Validation 35
3.5 Transcribed Utterance Segmentation 36
3.6 Syllabification Algorithm 36
3.7 Alignment Algorithm 38
4 Database Query 40
4.1 Terminology 40
4.2 Executing a Query 41
4.3 Creating a Query 42
4.4 An Illustrative Example 42
4.5 Additional Information 44
5 Future Projects 45
5.1 Interface for Acoustic Data 45
5.2 Extensions of Database Query Functionality 46
6 Discussion 47
References 48

Language Dynamics in the Framework of Complex Networks: A Case Study on Self-Organization of the Consonant Inventories 51
Animesh Mukherjee, Monojit Choudhury, Niloy Ganguly, and Anupam Basu
1 Introduction 51
2 Phonological Inventories: A Primer 53
3 Network Model of Consonant Inventories 56
3.1 Definition of PlaNet 56
3.2 Construction Methodology 57
4 Topological Properties of PlaNet 58
4.1 Degree Distribution of PlaNet 58
5 The Synthesis Model 61
6 Interpretation of the Synthesis Model 64
6.1 Mathematical Analysis of the Model 64
6.2 Linguistic Interpretation of the Model 66
7 Dynamics of the Language Families 68
8 Conclusion 71
Appendix 72
References 76

Part II Classifying Words and Mapping Them to Meanings

From Cues to Categories: A Computational Study of Children's Early Word Categorization 81
Fatemeh Torabi Asr, Afsaneh Fazly, and Zohreh Azimifar
1 Introduction 82
2 Related Work 83
2.1 Experimental and Corpus-Based Studies 84
2.2 Related Computational Models 84
3 Overview of This Study 85
4 Components of the Categorization Model 87
4.1 Categorization Algorithm 87
4.2 Cues Used in Categorization 88
5 Experimental Setup 89
5.1 Corpus 89
5.2 Feature Extraction 90
5.3 Model Parameters 91
6 Discovering Syntactic Categories 91
6.1 Evaluation Strategy 92
6.2 Novel Word Categorization 92
7 Word Categorization and Semantic Prediction 95
7.1 Semantic Feature Prediction 96
7.2 Simulation of the Brown Experiment 97
8 Conclusions and Future Directions 99
Appendix 102
References 102

In Learning Nouns and Adjectives Remembering Matters: A Cortical Model 105
Alessio Plebe, Vivian M. De la Cruz, and Marco Mazzone
1 Introduction 105
1.1 On Learning First Words 106
1.2 On Learning Nouns 108
1.3 On Learning Adjectives 109
1.4 Modeling Noun and Adjective Acquisition 110
2 Description of the Model 111
2.1 Basic Units of the Model 112
2.2 The Visual Pathway 113
2.3 Auditory Pathway 115
2.4 The Higher Cortical Map 116
3 Nouns and Adjectives Acquisition 118
3.1 Simulation of Intrinsic and Extrinsic Experience 119
3.2 Emergence of Organization in the Lower Maps 120
3.3 Representation of Nouns and Adjectives in Model PFC 120
3.4 Patterns of Connectivity of Nouns and Adjectives 123

4 Conclusions 125
References 125

Part III Learning Morphology and Syntax

Treebank Parsing and Knowledge of Language 133
Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick
1 Introduction: Treebank Parsing and Knowledge of Language 133
2 Experimental Methods 140
2.1 Parsing Systems Used 140
2.2 Training Data, Testing, and Evaluation 140
3 Case Study: Parsing Wh-Questions and QuestionBank 141
3.1 Augmenting the Training Data 143
4 Parsing and Tense: The Case of *Read* 146
5 Case Study: Parsing Passives by Linguistic Regularization 154
5.1 Passive Transformations: A Pilot Study 154
6 Parsing "Unnatural" Languages? 160
6.1 The Experimental Emulation 162
6.2 Training, Testing and Results 163
7 Discussion and Conclusions 167
References 169

Rethinking the Syntactic Burst in Young Children 173
Christophe Parisse
1 Introduction 173
2 Assumptions About Children's Behavior 174
3 A Testing Procedure in Three Steps 175
4 Analysis 1 177
5 Analysis 2 179
6 Analysis 3 182
6.1 Results and Discussion: Question 1 183
6.2 Results and Discussion: Question 2 184
7 Analysis 4 187
8 Discussion 188
Appendix 192
References 194

Part IV Linking Syntax to Semantics

Learning to Interpret Novel Noun-Noun Compounds: Evidence from Category Learning Experiments 199
Barry J. Devereux* and Fintan J. Costello
1 Introduction 200
2 An Exemplar-Based Account of Compound Interpretation 201

3 Overview of Experiments 202
4 Experiment 1 204
4.1 Method 205
4.2 Results 211
4.3 Discussion 215
5 Experiment 2 216
5.1 Method 217
5.2 Results 219
5.3 Modelling Relation Selection in Compound Interpretation 223
6 Experiment 3 225
6.1 Method 226
6.2 Results 228
6.3 Modelling Relation Selection 231
7 Conclusions 231
References 232

Child Acquisition of Multiword Verbs: A Computational Investigation ... 235
Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson
1 Introduction 236
2 Multiword Lexemes with Basic Verbs 238
3 Linguistic Properties and the Usage-Based Measures 239
3.1 Association of a Verb–Noun Pair 240
3.2 Semantic Properties of the Noun 241
3.3 Degree of Syntactic Fixedness 242
4 Evaluating the Statistical Measures 243
4.1 Experimental Setup 244
4.2 Measures in Combination: Clustering 245
4.3 Performance of the Individual Measures 246
5 Embedding the Measures into a Word Learning Model 247
5.1 The Original Word Learning Model 248
5.2 Learning the Verb–Noun Multiword Lexemes 249
5.3 Experiments on the Modified Word Learner 250
6 Conclusions 252
References 253

Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision 257
Michael Connor, Cynthia Fisher, and Dan Roth
1 Introduction 257
1.1 Addressing the Ambiguity of Sentences and Scenes: Semantic and Syntactic Bootstrapping 258
1.2 How Could Syntactic Bootstrapping Begin? 260
2 BabySRL and Related Computational Models 263
3 Model of Language Acquisition 267
3.1 CHILDES Training Data 268
3.2 Learning Model 269

4 Latent Training 274
4.1 Argument, Predicate and Role Classification 277
5 Experimental Evaluation 280
5.1 Ambiguous Semantic Feedback 282
6 Recovering Argument Knowledge 284
6.1 Bottom-Up Argument Identification 285
6.2 Integrating into Online Latent Classifier 288
7 Conclusion 290
References 293

Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model 297
Afra Alishahi and Suzanne Stevenson
1 Introduction 297
1.1 Verb Selectional Preferences 298
1.2 Related Computational Models 299
1.3 Our Approach 301
2 A Computational Model of Learning Verb Selectional Preferences 302
2.1 Learning as Bayesian Clustering 303
2.2 Probabilities of Semantic Properties 304
2.3 Predicting Semantic Profiles for Verbs 305
2.4 Verb-Argument Compatibility 306
3 Experimental Results 306
3.1 The Training Data 307
3.2 Formation of Semantic Profiles for Verbs 307
3.3 Evolution of Verb Semantic Profiles 310
3.4 Verb-Argument Plausibility Judgments 313
4 Discussion and Future Directions 314
References 315

Index 317

Contributors

Afra Alishahi Department of Communication and Information Studies, Tilburg University, Tilburg, The Netherlands

Zohreh Azimifar Computer Science and Engineering Department, Shiraz University, Shiraz, Iran

Anupam Basu Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Robert C. Berwick Massachusetts Institute of Technology, Cambridge, MA, USA

Rod Byrne Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Monojit Choudhury Microsoft Research India, Bangalore, Karnataka, India

Michael Connor Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA

Fintan J. Costello School of Computer Science and Informatics, University College Dublin, Dublin, Ireland

Vivian M. De la Cruz Department of Cognitive Science, University of Messina, Province of Messina, Italy

Barry J. Devereux Centre for Speech, Language and the Brain, Department of Psychology, University of Cambridge, Cambridge, UK

Afsaneh Fazly School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

Cynthia Fisher Department of Psychology, University of Illinois, Urbana-Champaign, IL, USA

Sandiway Fong University of Arizona Tuscon, AZ, USA

Niloy Ganguly Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Gregory J. Hedlund Department of Linguistics, Memorial University of Newfoundland, St. John's, NL, Canada

Dr. Anna Korhonen Department of Theoretical and Applied Linguistics (DTAL), University of Cambridge, Computer Laboratory, Cambridge, UK

Brian MacWhinney Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Igor Malioutov Massachusetts Institute of Technology, Cambridge, MA, USA

Marco Mazzone Laboratory of Cognitive Science, University of Catania, Catania, Italy

Animesh Mukherjee Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, West Bengal, India

Aida Nematzadeh Department of Computer Science, University of Toronto, Toronto, ON, Canada

Christophe Parisse MoDyCo-INSERM, CNRS, Paris Ouest Nanterre La Défense University, Nanterre cedex, France

Thierry Poibeau LaTTiCe-CNRS, Ecole Normale Supérieure and Université Sorbonne Nouvelle, Paris, France

Alessio Plebe Department Cognitive Science, University of Messina, Province of Messina, Italy

Yvan Rose Department of Linguistics, Memorial University of Newfoundland, St. John's, NL, Canada

Dan Roth Department of Computer Science, University of Illinois, Urbana-Champaign, IL, USA

Suzanne Stevenson Department of Computer Science, University of Toronto, Toronto, ON, Canada

Fatemeh Torabi Asr Computer Science and Engineering Department, Shiraz University, Shiraz, Iran

Aline Villavicencio Institute of Informatics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

Laboratory of Information and Decision Systems, Massachusetts Institute of Technology Cambridge, Cambridge, MA, USA

Todd Wareham Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Beracah Yankama Massachusetts Institute of Technology, Cambridge, MA, USA

Computational Modeling as a Methodology for Studying Human Language Learning

Thierry Poibeau, Aline Villavicencio, Anna Korhonen, and Afra Alishahi *

1 Overview

The nature and amount of information needed for learning a natural language, and the underlying mechanisms involved in this process, are the subject of much debate: how is the knowledge of language represented in the human brain? Is it possible to learn a language from usage data only, or is some sort of innate knowledge and/or bias needed to boost the process? Are different aspects of language learned in order? These are topics of interest to (psycho)linguists who study

*Excerpts of this chapter have been published in Alishahi, A. (2010), *Computational Modeling of Human Language Acquisition*, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers [2].

T. Poibeau (✉)
Laboratoire Langues, Textes, Traitements informatiques, Cognition, CNRS,
Ecole Normale Supérieure and Université Sorbonne Nouvelle, Paris, France
e-mail: thierry.poibeau@ens.fr

A. Villavicencio
Institute of Informatics, Federal University of Rio Grande do Sul, Av. Bento Gonçalves, 9500,
Porto Alegre, Brazil, 91501970
e-mail: alinev@gmail.com

A. Korhonen
University of Cambridge, Computer Laboratory, Cambridge, CB3 0FD, UK

Department of Theoretical and Applied Linguistics (DTAL), Cambridge, CB3 9DB, UK
e-mail: alk23@cam.ac.uk

A. Alishahi
Department of Communication and Information Studies, Tilburg University, Tilburg,
The Netherlands
e-mail: a.alishahi@uvt.nl

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_1, © Springer-Verlag Berlin Heidelberg 2013

human language acquisition, as well as to computational linguists who develop the knowledge sources necessary for large-scale natural language processing systems. Children are the ultimate subjects of any study of language learnability. They learn language with ease, in a short period of time and their acquired knowledge of language is flexible and robust.

Human language acquisition has been studied for centuries, but using computational modeling for such studies is a relatively recent trend. However, computational approaches to language learning have become increasingly popular, mainly due to advances in developing machine learning techniques, and the availability of large collections of experimental data on child language learning and child-adult interaction. Many of the existing computational models attempt to study the complex task of learning a language under cognitively plausible criteria (such as memory and processing limitations that humans face), and to explain the developmental stages observed in children. By simulating the process of child language learning, computational models can show us which linguistic representations are learnable from the input that children have access to in a reasonable amount of time, and which mechanisms yield the same patterns of behaviour that children exhibit during this process. In doing so, computational modeling provides insight into the plausible mechanisms involved in human language acquisition, and inspires the development of better language models and techniques.

The aim of this volume is to present a cross-section of recent research on the topic that draws on the relevance of computational techniques for understanding human language learning. These studies are inherently multidisciplinary, influenced by knowledge from fields such as Linguistics, Psycholinguistics and Biology, and the overview chapter starts with a discussion of some of the challenges faced, such as learnability constraints, data availability and cognitive plausibility. The strategies that have been adopted to deal with these problems build on recent advances in areas such as Natural Language Processing, Machine Learning, Artificial Intelligence and Complex Networks, as will be discussed in details in the chapters that compose this collection. Given the complex facets of language that need to be acquired, these investigations differ in terms of the particular language learning task that they target, and the overview chapter finishes with a contextualization of these contributions.

1.1 Theoretical Accounts of Language Modularity and Learnability

The study of human language acquisition pursues two important goals: first, to identify the processes and mechanisms involved in learning a language; and second, to detect common behavioural patterns in children during the course of language learning.

Languages are complex systems and learning one consists of many different aspects. Infants learn how to segment the speech signal that they receive as input, and they recognize the boundaries that distinguish each word in a sentence. They learn the phonology of their language, or the auditory building blocks which

form an utterance and the allowable combinations which form individual words. They assign a meaning to each word form by detecting the referent object or concept that the word refers to. They learn the regulations that govern form, such as how to change the singular form of a noun into a plural form, or the present tense of a verb into the past tense. They learn how to put words together to construct a well-formed utterance for expressing their intention. They learn how to interpret the relational meaning that each sentence represents and how to link different sentences together. On top of all these, they learn how to bring their knowledge of concept relations, context, social conventions and visual clues into this interpretation process.

A central question in the study of language is how different aspects of linguistic knowledge are acquired, organized and processed by the speakers of a language. The useful boundaries that break the language faculty into separate "modules" such as word segmentation, phonology, morphology, syntax, semantics and pragmatics, have been historically imposed to facilitate the study of each of these aspects in isolation. However, later psycholinguistic studies on language acquisition and processing suggest that the information relevant to these modules is not acquired in a temporally linear order, and that there is close interaction between these modules during both the acquisition and processing of language. In addition, many of the formalisms and processing techniques that have been proposed to handle a specific aspect may not be suitable for another.

The language modularity argument is part of a larger debate on the architecture of the brain, or the "modularity of mind." Proposals advocating a highly modular view rely extensively on the studies of Specific Language Impairments (SLI) which imply the isolation of language from other cognitive processes (e.g., [36]), whereas a highly interactive views refer to more recent studies on the interaction of language and other modalities such as vision or gesture at the process level (see Visual World Paradigm, [61]).

The modularity debate has been highly interleaved with the issue of nativism or language innateness. On the topic of language, the main point of interest has been whether humans are equipped with a highly sophisticated module for learning and using natural languages, consisting of task-specific procedures and representations, the "language faculty" [9, 11]. As complicated as it seems to master a language, children all around the world do it seemingly effortlessly and in a short period of time. They start uttering their first words around age 1. By the time they are 3–4-years old, they can use many words in various constructions, and can communicate fluently with other speakers of their native language. The efficiency with which children acquire language has raised speculations about whether they are born with some sort of innate knowledge which assists them in this process.

Human beings have an unparalleled skill for learning and using structurally complex languages for communication, and the learnability of natural languages has been one of the most controversial and widely discussed topics in the study of languages. The possibility of a genetic component that accounts for this unique ability of humans has been raised, but the extent and exact manifestation of this component is not clear. For instance, it has been argued that general learning and problem solving mechanisms are not enough to explain humans highly complex

communication skills, and some innate knowledge is also needed to account for their exceptional linguistic skills [13,53]. This hypothesis, known as the Innateness Hypothesis, states that human beings have an innate predisposition for learning languages, a task or domain specific knowledge, defined by their genetic code, and without having access to such innately specified linguistic knowledge a child cannot learn a language. Indeed, a nativist view of language learning states that natural languages are not learnable from the linguistic data that is typically available to children (Primary Linguistic Data, or PLD). The main argument in support of this view is the Argument from the Poverty of the Stimulus (APS; [9]), according to which PLD is both quantitatively and qualitatively too impoverished to allow for the acquisition of a natural language in its full structural complexity in a short period of time.

Of particular relevance to this discussion is the mathematical work of [28], who proved that a language learner cannot converge on the correct grammar from an infinitely large corpus without having access to substantial negative evidence. However, direct negative evidence (or corrective feedback from adult speakers of language) has been shown not to be a reliable source of information in child-directed data [44,45].[1] These findings have been viewed as compatible with nativist proposals for language acquisition such as that of a Universal Grammar (UG) [12], proposing that each infant is born with a innately specified representation of a grammar which determines the universal structure of a natural language. This universal grammar would be augmented by a set of parameters, which have to be adjusted over time to account for the particular language a child is exposed to.

In response to the nativist view of language learning, alternative representations of linguistic knowledge have been proposed, and various statistical mechanisms have been developed for learning these representations from usage data. A more empiricist view of language learning argues that a child does not have any innate prior knowledge about languages, and that languages can be learned using only general cognitive abilities which also underly other tasks (e.g. imitation, categorization and generalization [63, 64]) when these are applied to the sensory input to which a child is exposed. In an extreme version of empiricism, a child is like a *tabula rasa*, or a blank slate, when born, and all its language capabilities are learned from scratch from the environment [54].

Analyses of large collections of data on child-parent interactions have raised questions about the inadequacy of PLD [35, 54], arguing that child-directed data provides rich statistical cues about the abstract structures and regularities of language. In addition, recent psycholinguistic findings which hint at a 'bottom-up' process of child language acquisition have also questioned the top-down, parameter-setting approach advocated by the nativists. Advocates of this alternative view of language learning, also referred to as the *usage-based*, claim that children do not

[1] On the other hand, it has been suggested that the language learner can estimate the "typical" rate of generalization for each syntactic form, whose distribution serves as "indirect" negative evidence [15,42].

possess highly detailed linguistic knowledge at birth; instead they learn a language from the usage data they receive during the course of learning. Usage-based theories of language acquisition are motivated by experimental studies on language comprehension and generation in young children that suggest that children build their linguistic knowledge around individual items [1,3,38,39,63]. This view asserts that young children initially learn verbs and their arguments as lexical constructions and on an item-by-item basis, and only later begin to generalize the patterns they have learned from one verb to another. However, the details of the acquisition of these constructions and the constraints that govern their use are not clearly specified. Explicit models must be explored, both of the underlying mechanisms of learning these regularities, and of the use of the acquired knowledge.

In sum while nativism emphasises the role of *nature* as providing the required equipment, empiricism emphasises the role of *nurture* assuming that the environment is rich enough to provide a child with all the necessary evidence for language acquisition. Different proposals vary in terms of the extent in which they rely on language specific mechanisms and on general-purpose skills.

1.2 Investigations of Linguistic Hypotheses

One fundamental difficulty in research on language acquisition is that due to its characteristics it has to rely on indirect observation about the target processes. Apart from ethical considerations, the lack of non-invasive technology that is able to capture acquisition in action over time means that researchers can only assess different hypotheses indirectly, e.g. through diaries of child language, corpora of child-directed speech, or psycholinguistic data. As an strategy for probing human behaviour when learning and processing language, psycholinguistics provides a variety of experimental methodologies for studying specific behavioural patterns in controlled settings. Evidence concerning what humans (and children in particular) know about language and how they use it can be obtained using a variety of experimental techniques. Behavioural methods of studying language can be divided into two main groups: *offline* techniques, which aim at evaluating subjects' interpretation of a written or uttered sentence *after* the sentence is processed; and *online* techniques, which monitor the process of analyzing linguistic input *while* receiving the stimuli.

In offline studies, child language processing is examined in an experimental set-up using interactive methods in the form of *act-out scenarios* (when the experimenter describes an event and asks the child to act it out using a set of toys and objects), or *elicitation tasks* (when the child is persuaded to describe an event or action in the form of a natural language sentence). Preferential looking studies are another experimental approach conducted mostly on young children, where their preferences for certain objects or scene depictions is monitored while presenting them with linguistic stimuli.

In online methodologies, a variety of techniques are used (mostly on adult subjects) for identifying processing difficulties. A common technique in this category is measuring *reading times*. Many factors can affect reading times, therefore psycholinguistic studies use stimuli which are different in one aspect and similar in the others, and measure the reading time of each group of stimuli. Another technique that can be used on children as well as adult subjects is *eye-tracking*, where eye movements and fixations are spatially and temporally recorded while the subjects read a sentence on the screen. Using this technique, several reading time measures can be computed to evaluate processing difficulties at different points in the sentence. Also, anticipatory eye-movements can be analyzed to infer interpretations. Eye-tracking techniques have been employed in the *Visual World Paradigm* [61], where subjects' eye movements to visual stimuli are monitored as they listen to an unfolding utterance. Using this paradigm, the construction of online interpretation of a sentence and its mapping to the objects in the visual environment in real time can be studied.

More recently, *neuroscientific* methods have also been used for studying the processing of language in the brain. The most common approach is to measure *event-related potentials* (ERP) via electroencephalography (EEG): a stimulus is presented to the subject, while ERPs are measured through electrodes positioned on the scalp. Robust patterns have been observed in the change of ERPs as a response to linguistic stimuli. For example, when presented with a sentence with a semantic anomaly (e.g., *I like my coffee with cream and dog*), a negative deflection is usually observed 400 ms after the presentation of the stimuli. However, it is difficult to isolate the brain response to a particular stimulus, and it has been a challenge to derive a detailed account of language processing from such data. Functional Magnetic Resonance Imaging (fMRI) is another technique for measuring neural activity in the brain as a response to stimuli. Unlike EEG, fMRI cannot be used as an online measure, but it has higher spatial resolution and provides more accurate and reliable results.

In the majority of experimental studies of language, one aspect or property of the task or stimuli is selected and manipulated while other factors are held constant, and the effect of the manipulated condition is investigated among a large group of subjects. This approach allows researchers to isolate different language-related factors, and examine the significance of the impact that each factor might have on processing linguistic data. In such set-ups, it is only possible to manipulate the properties of the input data and the task in hand, and the learning or processing mechanisms that the subjects use for performing the task remain out of reach. Moreover, each subject has a history of learning and processing language which cannot be controlled or changed by the experimenter: all there is to control is a time-limited experimental session. Artificial languages are used to overcome any interference that the subjects' previous language-related experience might have on the outcome of the experiment. But the amount of the artificial input data that each subject can receive and process in these settings is very limited. These shortcomings call for an alternative approach for investigating the hypotheses regarding the acquisition and processing of natural languages.

2 Computational Models of Language Learning

Over the past decades, computational modeling has been used extensively as a powerful tool for in-depth investigation of existing theories of language acquisition and processing, and for proposing plausible learning mechanisms that can explain various observed patterns in child experimental data. The use of computational tools for studying language dates back to the onset of Artificial Intelligence. Early models mostly used logic rules for defining natural language grammars, and inference engines for learning those rules from input data. Over the last 20 years a rapid progress in the development of statistical machine learning techniques has resulted in the emergence of a wider range of computational models that are much more powerful and robust than their predecessors. As a result, computational modeling is now one of the main methodologies in the study of human cognitive processes, and in particular language.

Using computational tools for studying language requires a detailed specification of the properties of the input data that the language learner receives, and the mechanisms that are used for processing the data. This transparency offers many methodological advantages, such as:

- **Explicit definition of assumptions:** when implementing a computational model, every assumption, bias or constraint about the characteristics of the input data and the learning mechanism has to be specified. This property distinguishes a computational model from a linguistic theory, which normally deals with higher-level routines and does not delve into details, a fact that makes such theories hard to evaluate.
- **Control over input data:** unlike an experimental study on a human subject, the researcher has full control over all the input data that the model receives in its life time. This property allows for a precise analysis of the impact of the input on the behaviour of the model.
- **Control over experimental variables:** when running simulations of a model, the impact of every factor in the input or the learning process can be directly studied in the output (i.e., the behaviour) of the model. Therefore, various aspects of the learning mechanism can be modified and the behavioural patterns that these changes yield can be studied.
- **Choice of learning mechanisms:** the performance of two different mechanisms on the same data set can be compared against each other, something that is almost impossible to achieve in an experimental study on children.
- **Access to predictions of the model:** because of the convenience and the flexibility that computational modeling offers, novel situations or combinations of data can be simulated and their effect on the model can be investigated. This approach can lead to novel predictions about learning conditions which have not been previously studied.

Despite these advantages, computational modeling should not be viewed as a substitute but rather as a complement for theoretical or empirical studies of language. One should be cautious when interpreting the outcome of a computational

model: if carefully designed and evaluated, computational models can show what type of linguistic knowledge is learnable from what input data. Also, they can demonstrate that certain learning mechanisms result in behavioural patterns that are more in line with those of children. In other words, computational modeling can give us insight into which representations and processes are more plausible in light of the experimental findings on child language acquisition. They can be viewed as the testing grounds for different theories and can provide information about the conditions under which these would succeed in a given task. However, even the most successful computational models can hardly prove that humans exploit a certain strategy or technique when learning a language. Cognitive scientists can use the outcome of computational modeling as evidence on what is possible and what is plausible, and verify the suggestions and predictions made by models through further experimental and neurological studies.

2.1 What to Expect from a Model

Traditionally, linguistic studies of language have been focused on representational frameworks which can precisely and parsimoniously formalize a natural language according to how adult speakers of that language use it. In this approach, the focus is on the end product of the acquisition process, and not on the process itself. On the other hand, psycholinguistic studies mainly emphasize the process of learning and using a language rather than the language itself [14]. This dual approach is also reflected in the modeling of language acquisition. One modeling strategy is to demonstrate the feasibility of extracting an optimal structure from a given linguistic input (e.g., a grammar from a text corpus, or a phonetic or lexical segmentation from a large stream of speech signals), aiming at compatibility of the results with a target. An alternative strategy is to focus on developmental compatibility and replicate the stages that children go through while learning a specific aspect of language, such as vocabulary growth in word learning or the U-shaped generalization curve in the acquisition of verb argument structure. Therefore, given the priorities of each of these strategies it is important to evaluate a model in the context that it is developed in, and with respect to the goals that it is aiming at.

Another critical point when assessing a model is to identify the fundamental assumptions that the model is based on. When developing a model for computational simulation of a process, all the details of the process must be implemented, and no trivial aspect of the representational framework or the procedure can be left unspecified. However, many of these details are of secondary importance to the process that the model aims to study. It is of utter importance for the developers of a computational model to clearly specify which theoretical assumptions about the implemented model or the characteristics of the input data are fundamental, and which implementation decisions are arbitrary. Moreover, they must show that the overall performance of the model does not crucially depend on these trivial decisions.

Finally, the level of processing targeted by a model must also be taken into account. One of the first (and most influential) categorizations of cognitive models was proposed by Marr [47], who identifies three levels of describing a cognitive process. First is the *Computational* level, which identifies what knowledge is computed during the process. This is the highest level a model can aim for: the focus is on what is needed or produced during the cognitive process under study, abstracting from any learning or processing algorithm that is used for computing or applying this knowledge. Next comes the *Algorithmic* level, which specifies how computation takes place. At this level, the focus is on the mechanisms involved in the computational process. Finally there is the *Implementation* level, which simulates how the algorithms are actually realized in brain. At this level, every implementational detail is a vital component of the model. It is important on the modelers' side to specify, and on the evaluators' side to take into account, the intended level of the model to be assessed. If the simulation of a model aimed at a computational level of describing a process results in a behavioural pattern that is inconsistent with that of children, it might be due to an inappropriate choice of algorithm or other implementational details, and not because the specification of the proposed computation itself is flawed.

One important constraint when developing a cognitive model is *cognitive plausibility*. In the field of natural language processing, many automatic techniques have been developed over the years for extracting various types of linguistic knowledge from large collections of text and speech, and for applying such knowledge to different tasks. In this line of research, the main goal is to perform the task at hand as efficiently and accurately as possible. Therefore, any implementation decision that results in better performance is desired. For instance, to induce wide coverage grammars from corpora, supervised learning methods based on annotated data such as the Penn Treebank have been usually employed, with grammars that tend to be less than or equal to context free grammars in expressive power and which may not be linguistically adequate to capture human grammar [60]. However, cognitive models of language learning and processing are not motivated by improving performance on a certain task. Instead, they are aiming at simulating and explaining how humans perform that task. Such models have to conform to the limitations that humans are subject to.

A model which attempts to simulate a cognitive process has to make realistic assumptions about the properties of the input data that are available to children during that process. For example, a model of syntax acquisition cannot assume that children are being corrected when producing an ungrammatical sentence, since various analyses of child-directed data have shown that such information is not consistently provided to them [44]. Also, when modeling any aspect of child language acquisition, it cannot be assumed that children receive *clean* input data, since the data almost always contain a high level of noise and ambiguity. Sometimes it is inevitable to make simplifying assumptions about the structure of data in order to keep calculations feasible or to focus on one specific aspect of learning. However, if a model makes obviously false assumptions about the input, any finding by such a model might not be generalizable to realistic situations.

Also, a cognitive model must draw on language-independent strategies. Children around the world learn a variety of languages with drastically different characteristics, such as their sound system or structure. It is highly implausible to assume that children use different learning mechanisms for learning different languages. Thus a model of language learning must avoid any language-specific assumptions or learning strategies. For example, a model of learning syntax which assumes a rigid word order cannot be extended to families of languages with a more relaxed word order.

Finally, cognitive models must conform to the memory and processing limitations of humans. The architecture of the human brain and its processing capacities and memory resources are very different from those of the existing computational systems. Thus many of the machine learning techniques that are developed for applying on large-scale data sets are not suitable for modeling human language processing. For example, it is unlikely that children can remember every instance of usage of a particular word or every sentence that they have heard during their lifetime in order to learn something about the properties of language. This limits the scope of the techniques and algorithms that can be used in cognitive modeling. One of the by-products of human memory and processing limitations is that language must be learned in an incremental fashion. Every piece of input data is processed when received, and the knowledge of language is built and updated gradually. This is in contrast with many machine learning techniques which process large bodies of input at once (usually through iterative processing of data) and induce an optimum solution (e.g., a grammar) which formalizes the whole data set precisely and parsimoniously.

Although a cognitive model of language is often expected to provide a cognitively plausible explanation for a process, it is the intended description level of the model which determines the importance of various plausibility criteria. For example for a model at the computational level, making realistic assumptions about the characteristics of the input data is crucial. However, conforming to processing limitations (such as incrementality) in the implementation of the model is of secondary importance, since the model is not making any claims about the actual algorithm used for the proposed computation.

2.2 Modeling Frameworks

The first generation of models of language were influenced by early artificial intelligence techniques, including the logic-based inference techniques which were widespread in 1960s. Symbolic modeling often refers to an explicit formalization of the representation and processing of language through a symbol processing system. In this approach, linguistic knowledge is represented as a set of symbols and their propositional relations. Processing and updating this knowledge takes place through general rules or schemas, restricted by a set of constraints. Each rule might be

augmented by a list of exceptions, i.e., tokens or instances for which the rule is not applicable. The syntactic structure of a language is typically modeled as a rule-based grammar, whereas the knowledge of semantics is modeled through schemas and scripts for representing simple facts and processes. These representations are often augmented by a set of logical rules for combining them and constructing larger units which represent more complex events and scenarios.

Following the Chomskian linguistics tradition, symbolic models of language assume that a language is represented as an abstract rule-based grammar which specifies all (and only) valid sentences, based on judgements of linguistic acceptability [12]. In this view, language processing is governed by internally specified principles and rules, and ambiguities are resolved using structural features of parse trees (e.g., the principle of minimal attachment; [24]). The influence of lexical information on parsing and disambiguation is often overlooked by these theories. Language acquisition, on the other hand, has been mainly modeled through *trigger-based* models, where the parameters associated with a pre-specified grammar are set to account for the linguistic input data (e.g., [26]).

Symbolic models of language are often transparent with respect to their linguistic basis, and are computationally well-understood. However, typical symbolic models do not account for the role of *experience* (or the statistical properties of the input data) on behaviour and are not robust against noise and ambiguity.

Connectionist models of cognitive processes [48] emerged during 1980s as an alternative to symbolic models. The architectural similarities between the connectionist models and the human brain on a superficial level, and their capacity for distributional representation and parallel processing of knowledge made them an appealing choice for modeling human language acquisition and processing. The idea of connectionist models is based on simple neural processing in brain. Each connectionist model (or *artificial neural network*) consists of many simple processing units (or *neurons*), usually organized in layers, which are heavily interconnected through weighted links. Each neuron can receive many input signals, process them and pass the resulting information to other neurons. Linguistic knowledge is represented as distributed activation patterns over many neurons and the strength of the connections between them. Learning takes place when connection weights between neurons change over time to improve the performance of the model in a certain task, and to reduce the overall error rate. A cognitive process is modeled by a large number of neurons performing these basic computations in parallel.

Various versions of artificial neural networks have been proposed which vary in the neuron activation function, the architecture of the network, and the training regime. For modeling language learning, multi-layered, feed-forward networks have been most commonly used. These networks consist of several neurons, arranged in layers. The input and output of the cognitive process under study are represented as numerical vectors, whose dimensions correspond to input units. Such models are normally trained in a supervised fashion: the model produces an output for a given input pattern, and the connection weights are adjusted based on the difference between the produced and the expected output.

Connectionist models have received enormous attention from the cognitive science community due to the learning flexibility they offer compared to symbolic models (e.g., [20,40]), and because they suggest that general knowledge of language can be learned from instances of usage. However, these models often cannot easily scale up to naturalistic data. Moreover, the knowledge they acquire is not transparent, and is therefore hard to interpret and evaluate.

The relatively recent development of machine learning techniques for processing language motivated many researchers to use these methods as an alternative modeling paradigm. Probabilistic modeling allows for combining the descriptive power and transparency of symbolic models with the flexibility and experience-based properties of the connectionist models. Probabilities are an essential tool for reasoning under uncertainty. In the context of studying language acquisition, probabilistic modeling has been widely used as an appropriate framework for developing experience-based models which draw on previous exposure to language, and at the same time provide a transparent and easy to interpret linguistic basis. Probabilistic modeling views human language processing as a rational process, where various pieces of evidence are weighted and combined through a principled algorithm to form hypotheses that explain data in an optimal way. This view assumes that a natural language can be represented as a probabilistic model which underlies sentence production and comprehension. Language acquisition thus involves constructing this probabilistic model from input data.

Many probabilistic models of language are essentially an augmented version of their symbolic counterparts, where each rule or schema is associated with a weight (or probability). For example, Probabilistic Context Free Grammars (PCFG) use a symbolic representation of the syntactic knowledge (CFG), but they also calculate a probability for each grammar rule depending on the number of times that rule has appeared in the input [32]. However, an alternative (and more radical) probabilistic view proposes language represented as a bottom-up, graded mapping between the surface form and the underlying structure, which is gradually learned from exposure to input data (e.g., [16, 64]).

The acquisition of linguistic knowledge can be formulated as an induction process, where the most likely underlying structures are selected based on the observed linguistic evidence. The basic idea behind this process is to break down complex probabilities into those that are easier to compute, often using Bayes' rule. A family of probabilistic models, generally referred to as Bayesian models, have gained popularity over the past decade. In the context of grammar learning, Bayesian methods specify a framework for integrating the prior information about the grammatical structures and the likelihood of the observed word strings associated with each structure, to infer the most probable grammatical structure from a sentence. The prior probabilities are often used for embedding underlying assumptions about the hypothesis space and for seemlessly integrating biases and constraints into the system. It has been argued that prior information (specifically the prior structure over Bayesian networks) is crucial to support learning [62]. The positioning of these models in relation to nativism due to the nature of the prior information adopted remains to be determined. As [51] discuss, there seems to be an agreement along the

empiricist-nativist continuum that there must be some innate constraints. Different proposals vary as to which constraints are adopted and how strong and domain-specific they are, given the empiricist ideal of a bottom-up, data-driven learning.

In addition to the probabilistic frameworks that are specifically developed for representing and processing linguistic knowledge, many recent computational models heavily rely on general-purpose statistical machine learning tools and techniques. A variety of such methods have been successfully exploited in more practical natural language processing applications. The efficiency of these methods has motivated their use in modeling human language acquisition and processing, in particular for the purpose of extracting abstract and high-level knowledge from large collections of data. In this context, one approach from Information Theory that has been adopted in various computational language acquisition tasks is the Minimum Description Length (MDL) Principle. MDL is an algorithmic paradigm for evaluating the hypothesis space, based on Occam's Razor, in which the best hypothesis for a given set of data is the one that leads to the best compression of the data [55]. The idea is that MDL can be used to order the hypothesis space according to how compact the hypotheses are and how well they generate the data [37]. MDL has proved to be a powerful tool in many language related tasks, such as word segmentation (e.g., [4, 17]), grammar learning (e.g., [19, 30, 33, 66]) and learnability assessment (e.g., [31]).

Probabilistic models in general are robust against noise, and are a powerful tool for handling ambiguities. A range of statistical and probabilistic techniques have been efficiently employed over the last couple of decades to modeling various aspects of language acquisition and use, some examples of which can be seen in the papers in this collection. However, some suggest that probabilistic methods must be viewed as a framework for building and evaluating theories of language acquisition, rather than as embodying any particular theoretical viewpoint [8].

2.3 Research Methods

As a response to the nativist claims that some aspects of language (mainly syntax) are not learnable solely from input data, a group of computational models have been proposed to challenge this view and investigate to what extent extracting a grammatical representation of language from a large corpus is in fact possible. These models are not considered as typical cognitive models, since most of them are not concerned with how humans learn language. Instead, their goal is to show that the Primary Linguistic Data is rich enough for an (often statistical) machine learning technique to extract a grammar from it with high precision and without embedding any innate knowledge of grammar into the system. On the other hand, a typical cognitive model cannot be solely evaluated based on its accuracy in performing a task. The behaviour of the model must be compared against observed human

behaviour, the errors made by humans must be replicated and explained, and the result must be linguistically and psychologically motivated. Therefore, evaluation of cognitive models depends highly on the experimental studies of language.

We need to compare the knowledge of a cognitive model to that of humans in a particular domain. But there is no direct way to figure out what humans *know* about language. Instead, their knowledge of language can only be estimated or evaluated through how they *use* it in language processing and understanding, as well as in language production. Analysis of child production data provides valuable cues about the trajectory of their learning the language. Many developmental patterns are revealed through studying the complexity of the utterances that children produce, the errors that they make and the timeline of their recovery from these errors. On the other hand, comprehension experiments reveal information about knowledge sources that children exploit, their biases towards linguistic and non-linguistic cues, and their awareness of the association between certain utterances and events.

Earlier studies of child language acquisition were based on sporadic records of interaction with children, or isolated utterances produced by children which researchers individually recorded. But recent decades have seen a significant growth in the variety and quantity of resources for studying language, and a collective attempt from the computational linguistics and cognitive science communities to use standard formats for the expansion of these resources.

The most well-known and widely used database of transcriptions of dialogues between children and their caregivers is CHILDES [20], a collection of corpora containing recorded interactions of adults with children of different age and language groups and from different social classes. Transcriptions are morphologically annotated and mostly follow a (semi-)standard format, and occasionally, some semantic information about the concurrent events is added to the conversation (e.g., what objects are in the scene or what the mother points to). The English portion of CHILDES has been annotated with dependency-based syntactic relations [59]. Many of the databases in CHILDES also contain audio or video recordings of the interaction sessions, but these recordings are mostly unannotated.

Some of the audio and video recordings in CHILDES have been annotated by individual research groups for specific purposes. For example, [68] and [23] use video clips of mother-infant interactions from CHILDES, and manually label the visible objects when each utterance is uttered, as well as the objects of joint attention in each scene. Other social cues such as gaze and gesture are also marked. A more systematic approach is taken by the TalkBank project, which is accumulating the speech corpus of children with multimodal annotation [43]. Other researchers have collected smaller collections of annotated videos from children. One such example is the recording of adults reading story books to 18 month old infants, annotated to identify the physical objects and the spoken words in each frame in the video [69]. Another example is a set of videos of a human operator enacting visual scenes with toy blocks, while verbally describing them [18]. These resources are sparse, and the annotation scheme or the focus of annotation is rather arbitrarily chosen by the researchers who developed them. Another massive collection of data has

been recently gathered by Roy [56]. Roy has recorded his son's development at home by gathering approximately 10 h of high fidelity audio and video on a daily basis from birth to age 3. However, the resulting corpus is not structured. These collections are hard to use without some sort of preprocessing or manual annotation. Nevertheless, they are complementary to the textual data from corpora which lack any semantic information.

Ever since the availability of resources like CHILDES [20], both child-directed (utterances by parents and other adults aimed at children) and child-produced data have been extensively examined. Analyses of child-directed data in particular have mainly focused on the grammaticality of the data, its statistical properties (e.g., [31]), and the availability of various cues and constructions (e.g., [29]). Such analyses have provided valuable information about what children have access to. For example, child-directed data has been shown to be highly grammatical (e.g., [5]), and sufficiently rich with statistical information necessary for various tasks (e.g., the induction of lexical categories [49]).

Utterances produced by children, on the other hand, have been analyzed with a different goal in mind: to identify the developmental stages that children go through in the course of learning a language, and to detect common behavioural patterns among children from different backgrounds. The parameters usually examined in child-produced data are: (a) the size of the vocabulary that they use; (b) the length of the sentences that they produce; (c) the complexity of these sentences (which syntactic constructions they use); (d) the wide-spread errors that they make and the type of these errors; and (e) how each of these factors changes as the child ages. Also, differences between each of these factors have been studied in children of different genders, nationalities and social classes. Such studies have yielded substantial evidence about children's learning curves in different tasks (e.g., word learning or argument structure acquisition).

Besides the more direct use of adult-child interaction data, properties of the data are also used in evaluating computational models of language. Statistical properties of child-directed data (average sentence length, distributional properties of words, etc.) are normally used as standard when selecting or artificially creating input for many computational models (e.g., [7, 31, 51, 66]). Additionally, several models have attempted to simulate or explain the patterns observed in child-produced data.

In addition to these child-focused collections, there are several large corpora of adult-generated text and speech. These corpora, such as the Brown corpus [22], the Switchboard corpus [27], and the British National Corpus (BNC; [6, 34]) contain large amounts of data, and are representative of language used by a large number of speakers of a language (mostly English) in different domains and some of these corpora are entirely or partially annotated with part of speech tags or parsed (e.g., [46]). Although these corpora are normally used as input data for models of grammar induction, they have also been used as basis for comparative and even complementary analyses to those reported using child-related data (e.g., [31]).

3 Impact of Computational Modeling on the Study of Language

Advances in machine learning and knowledge representation techniques have led to the development of powerful computational systems for the acquisition and processing of language. Concurrently, various experimental methodologies have been used to examine children's knowledge of different aspects of language. Empirical studies of child language have revealed important cues about what children know about language, and how they use this knowledge for understanding and generating natural language sentences. In addition, large collections of child-directed and child-produced data have been gathered by researchers. These findings and resources have facilitated the development of computational models of language. Less frequently, experiments have been designed to assess the predictions of some computational models on a particular learning process. Computational cognitive modeling is a new and rapidly developing field. During its short life span, it has been extensively beneficial to cognitive science in general, and the study of natural language acquisition and use in particular.

One of the main impacts of computational models of language acquisition has been to emphasize the importance of probabilistic knowledge and information theoretic methods in learning and processing language. The role of statistical methods in language acquisition for long in the sidelines has been gaining prominence in recent years [2]. The undeniable success of statistical techniques in processing linguistic data for more applied NLP tasks has provided strong evidence for their impact in human language acquisition [8]. On the other hand, shallow probabilistic techniques which are not linguistically motivated can only go so far. For example, pure distributional models have been generally unsuccessful in accounting for learning a natural language in realistic scenarios. Fifty years after the development of the first computational models of language, hybrid modeling approaches that integrate deep structures with probabilistic inference seem to be the most promising direction for future research.

Developing computational algorithms that capture the complex structure of natural languages in a linguistically and psychologically motivated way is still an open problem. Computational studies of language combine research in linguistics, psychology and computer science. Because it is a young field of a highly interdisciplinary nature, the research methods employed by scholars are inevitably varied and non-standard. This is an unfortunate situation: it is often difficult to compare different models and analyze and compare their findings due to incompatible resources and evaluation techniques they employ. It is vital for the community to share resources and data collections, to develop unified schemes for annotation and information exchange, and to converge on standards for model comparison and evaluation.

When it comes to comparing and evaluating computational models, there is even less agreement among researchers in this field. The majority of algorithms used for simulating language acquisition are unsupervised, mainly because it is highly

unrealistic to assume that children receive input data which is marked with the kind of linguistic knowledge they are supposed to learn. As a consequence, there is no gold standard for evaluating the outcome of these unsupervised models and the success of their results and contributions may be difficult to assess. Furthermore, the underlying representation of the linguistic knowledge in human brain is unknown; therefore, the knowledge of language that a model acquires cannot be evaluated on its own. Many models apply their acquired knowledge on different tasks, but such tasks are often chosen arbitrarily. With computational modeling becoming more widespread, it is extremely important for the community to converge on standard evaluation tasks and techniques in each domain that can be used for rigorous evaluation of the methodology and replicability of the results, as in the more traditional disciplines that influence the field.

4 This Collection

4.1 Methods and Tools for Investigating Phonetics and Phonology

Child language corpora are essential for research on language acquisition, yet prohibitively expensive to build. The study of child language acquisition has made great progress in recent years thanks to the availability of shared corpora and tools among researchers. The CHILDES database includes a large number of corpora for growing number of languages like Danish, Portuguese, German, Russian and Cantonese among others. Moreover, the tools associated with CHILDES (e.g. CLAN) enable easy access to the information provided in corpora allowing complex searches that combine different levels of information. However, the majority of the tools associated with CHILDES focus on morphology, syntax and semantics [50,58], with a lack of tools for phonetics and phonology. With the development of Phon this is no longer the case. The chapter *Phon 1.4: A Computational Basis for Phonological Database Elaboration and Model Testing* of this collection by Yvan Rose, Gregory J. Hedlund, Rod Byrne, Todd Wareham, and Brian MacWhinney introduces version 1.4 of Phon – an open-source software program designed for the transcription, coding, and analysis of phonological corpora. Phon 1.4. is a versatile program capable of supporting multimedia data linkage, utterance segmentation, multiple-blind transcription, transcription validation, syllabification, and the alignment of target and actual forms. The system is available with a graphical interface and is used by PhonBank, a database project that seeks to broaden the scope of CHILDES into phonological development and disorders.

A framework for phonetic investigations is also the topic of the chapter *Self-Organization of the Consonant Inventories in the Framework of Complex Networks*, by Animesh Mukherjee, Monojit Choudhury, Niloy Ganguly and Anupam Basu. It introduces a computational framework for representing, analysing and synthesising consonant inventories of the world's languages. The framework is capable of

integrating the full variety in consonants as well as languages. It is essentially a complex network with two sets (or partitions) of nodes: one for consonants and the other one for languages. The primary objective is to provide the means to systematically analyse and synthesise the structure of the Phoneme-Language Network (PlaNet) and thereby, explain the distribution with which consonants occur across languages.

4.2 Classifying Words and Mapping Them to Meanings

The task for learning the meanings of words can be viewed as children learning the association between a word form and a concept after hearing repeated instances of the word form used in reference to the concept. Despite its misleading apparent simplicity, there are many challenges to this task. First, few words are used in isolation. Children usually hear words in a sentential context. Secondly, a sentence can potentially refer to many different aspects of a scene, as a typical learning situation usually involves a large number of objects. For a learner who does not know the meanings of words yet, it can be a real challenge to figure out the exact aspect (or the relational meaning) that the sentence conveys. Thirdly, child-directed speech has been shown to contain a substantial level of noise and ambiguity. Therefore learning the correct mapping between each word and its meaning is a complex process that needs to be accounted for by a detailed model.

In addition to the acquisition of word meanings, psycholinguistic studies suggest that early on children acquire robust knowledge of some of the abstract lexical categories such as nouns and determiners. For example, [25] show that 2-year-olds treat novel words which do not follow a determiner (e.g., *Look! This is Zag!*) as a proper name which refers to an individual. In contrast, they tend to interpret novel words which do follow a determiner (e.g., *Look! This is a zag!*) as a mass noun. However, learning lexical categories takes place gradually, and not all categories are learned at the same time. For example, [65] show that 2-year-olds are more productive with nouns than with verbs, in that they use novel nouns more frequently and in more diverse contexts. How children gain knowledge of syntactic categories is an open question. Children's grouping of words into categories might be based on various cues, including the phonological and morphological properties of a word, distributional information about its surrounding context, and its semantic features.

The chapter of Fatemeh Torabi Asr, Afsaneh Fazly, and Zohreh Azimifar's *From cues to categories: A computational study of children's early word categorization* focuses on the acquisition of word categories. The authors investigate the types of information children require in order to learn these categories. The paper proposes a computational model which is capable of acquiring categories from distributional, morphological, phonological, and semantic properties of words. The results show that syntax plays an important role in learning word meaning (and vice versa). Additionally, the proposed model can predict the semantic class of a word (e.g., action or object) by drawing on the learned knowledge of the word's syntactic category.

In the chapter *In learning nouns and adjectives remembering matters: a cortical model*, Alessio Plebe, Vivian De la Cruz and Marco Mazzone investigate different artificial models which can be used to mimic the acquisition of mapping of words to meanings. The authors use a hierarchy of artificial cortical maps to develop models of artificial learners that are subsequently trained to recognize objects, their referents, and the adjectives pertaining to their color. In doing so, they address several fundamental issues such as noun acquisition as well as adjective acquisition (which is known to be a much more difficult task). The relation between nouns and adjectives also plays a role in their meaning, so the model accounts for an embryonic syntax. Results reported in the chapter explain various developmental patterns followed by children in acquiring nouns and adjectives, by perceptually driven associational learning processes at the synaptic level.

4.3 Learning Morphology and Syntax

Learning the categories or meanings of words is not enough for successful communication: a language learner has to also master the regularities that govern word forms, and the acceptable combinations of words into sentences. Natural languages are highly regular in their morphological and syntactic structure. Regularities in syntax, such as the position of the subject and object in relation to the verb, can provide important information for the learner about the structure of the language (e.g. the SVO order in English, and SOV in Japanese). Nevertheless, in each language there are words which do not conform to such general patterns, and one well known case is that of exceptions to the English past tense verb formation [45, 57, 67] with *ed* suffix (e.g. *receive/received* vs. *ring/rang/*ringed*). The challenge in learning morphology and syntax is how to grasp the abstract regularities that govern form, as well as the idiosyncratic properties of individual words and constructions based on potentially poor stimulus and/or no consistent negative evidence.

One well known example in the discussion on the poverty of the stimulus centers around the knowledge of structure dependency in question inversion [10], and whether the relevant data provides sufficient information to guarantee successful acquisition. For instance, for native speakers in general the following two sentences are closely related:

1. *The company has bought his shares.*
2. *Has the company bought his shares?*

but for learners they provide a though challenge, as a learner has to identify the relation between the declarative sentence and the corresponding derived interrogative form without overgenerating to possible but ungrammatical, unattested or unnatural forms (e.g. **Bought the company has his shares?* and *?Has bought the company his shares?*).

Related to this discussion is the chapter by Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick in the chapter *Treebank Parsing and Knowledge of Language* who examine some complex linguistic constructions with non-local dependencies that are also challenging for computational models, such as tense marking, wh-questions and passives in English, assessing gaps in the *knowledge of language* acquired from large corpora. They investigate if grammars acquired by statistical parsers can successfully account for a full knowledge of language, verifying in which cases poor performance may be due to data sparsity, and in which it might arise from the underlying grammatical frameworks. They propose a general approach which incorporates linguistic knowledge by means of *regularizations* that canonicalize predicate-argument structure, which results in statistically significant improvements in parser performance. The results obtained indicate the contributions of distributional and linguistic properties of data needed for a successful account of language, and where insights from language acquisition can positively inform statistical parsing development.

In other cases, the data may provide enough information for a learning mechanism to obtain results compatible with syntactic evolution in language acquisition. Christophe Parisses's chapter, *Rethinking the syntactic burst in young children*, focuses on the speed and correctness of child language acquisition at the point of "syntactic burst" which usually occurs around age 2–3. The author shows that recent theories based on general cognitive capabilities such as perception, memory or analogy processing do not sufficiently explain the syntactic burst. He then proposes a testing procedure to demonstrate that the acquisition of usage-based and fixed-form patterns can account for the syntactic evolution in language acquisition. The patterns are applied to the Manchester corpus taken from the CHILDES database. The author shows that simple mechanisms explain language development until age 3 and that complex linguistic mastery does not need to be available early in the course of language development.

4.4 Linking Syntax to Semantics

Experimental child studies have shown that children are sensitive to associations between syntactic forms and semantic interpretations from an early age, and that they use these associations to produce novel utterances [3, 41, 52]. Children's learning of form-meaning associations is not well understood. Specifically, it is not clear how children learn the item-specific and general associations between meaning and syntactic constructions.

One aspect of language that provides a rich testbed for studying form-meaning associations is the argument structure of verbs. The argument structure of a verb determines the semantic relation a verb has with its arguments and how the arguments are mapped onto valid syntactic expressions. This complex structure exhibits both general patterns across semantically similar verbs, as well as more idiosyncratic mappings of verbal arguments to syntactic forms. This is particularly

acute in the case of multiword expressions, which vary along a continuum from literal to more idiomatic cases, like the phrasal verbs *carry up* and *come up (with an idea)* (as *propose an idea*), respectively, and from more to less productive expressions (e.g. *carry up/down* vs. *come up/?down*).

In addition to regularities at the level of argument structure, research on child language has revealed strong associations between general semantic roles such as Agent and Destination and syntactic positions such as Subject and Prepositional Object (e.g., [21], and related work). Despite the extensive use of semantic roles in various linguistic theories, there is little consensus on the nature of these roles. Moreover, researchers do not agree on how children learn general roles and their association with grammatical functions.

The first chapter on this topic, *Learning to interpret novel noun-noun compounds: Evidence from category learning experiments* by Barry J. Devereux and Fintan J. Costello, focuses on the analysis of one type of multiword expressions: nominal compounds. The interpretation of noun-noun compounds is known to be challenging and difficult to predict since these constructions are highly ambiguous. Two approaches have been proposed for the interpretation of noun-noun compounds: one which assumes that people make use of distributional information about the linguistic behaviour of words and how they tend to combine as noun-noun phrases; another which assumes that people activate and integrate information about the two constituent concepts' features to produce interpretations. Devereux and Costello propose a model that combines these two approaches. They present an exemplar-based model of the semantics of relations which captures these aspects of relation meaning, and show how it can predict experimental participants' interpretations of novel noun-noun compounds.

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson, in the chapter *Child Acquisition of Multiword Verbs: A Computational Investigation*, address the question of the acquisition of multiword expression by children. They show that multiword expressions have received far less attention than simple words in child language studies. However, in natural language processing, there is a long research tradition on models for the recognition and analysis of idiosyncratic expressions. The authors explore whether this computational work on multiword lexemes could be extended in a natural way to the domain of child language acquisition where an informative cognitive model must take into account two issues: what kind of data the child is exposed to, and what kind of processing of that data is cognitively plausible for a child. They also present a word learning model that uses this information to learn associations between meanings and sequences of words.

In *Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision* Michael Connor, Cynthia Fisher and Dan Roth investigate the problem of assigning semantic roles to sentence constituents, where a learner needs to parse a sentence, find possible arguments for predicates, and assign them adequate semantic roles. They look at possible starting points for a learner using a computational model, Latent BabySRL, which learns semantic role classification from child-directed speech. They found that even before acquiring any specific verb knowledge this model is able to begin comprehending simple semantics in a plausible setup when initialized with a small amount of knowledge about nouns and some biases.

In *Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model*, Afra Alishahi and Suzanne Stevenson present a cognitive model for inducing verb selectional preferences from individual verb usages. The selectional preferences for each verbal argument are represented as a probability distribution over the set of semantic properties that the argument can possess, i.e. a semantic profile. The semantic profiles yield verb-specific conceptualizations of the arguments associated with a syntactic position. The proposed model can learn appropriate verb profiles from a small noisy training data, and can use them to simulate human plausibility judgments and to analyse implicit object alternation.

5 Concluding Remarks

These chapters present a cross-section of the research on computational language acquisition, and investigate linguistic and distributional characteristics of the learning environment for different linguistic aspects, adopting a variety of learning frameworks. Computational investigations like these can contribute to research on human language acquisition, challenging the extent to which innate assumptions need to be specified in these models, and how successful they are in each of the specific tasks, providing valuable insights into learnability aspects of the data, the learning environment and the specific frameworks adopted. This is a new and fast growing multidisciplinary field that has yet much to achieve, evolving along with its foundational areas.

References

1. Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language, 26*, 339–356.
2. Alishahi, A. (2010). *Computational modeling of human language acquisition* (Synthesis lectures on human language technologies). San Rafael: Morgan & Claypool Publishers.
3. Bowerman, M. (1982). Evaluating competing linguistic models with language acquisition data: Implications of developmental errors with causative verbs. *Quaderni di semantica, 3*, 5–66.
4. Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*(1–2), 93–125.
5. Broen, P. A. (1972). *The verbal environment of the language-learning child*. Washington: American Speech and Hearing Association.
6. Burnard, L. (2000). *Users reference guide for the British National Corpus* (Technical Report). Oxford University Computing Services.
7. Buttery, P., & Korhonen, A. (2007). I will shoot your shopping down and you can shoot all my tins: Automatic lexical acquisition from the CHILDES database. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* (pp. 33–40). Prague: Association for Computational Linguistics.
8. Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Science, 10*(7), 335–344.
9. Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

10. Chomsky, N. (1975). *The logical structure of linguistic theory*. New York: Plenum press.
11. Chomsky, N. (1980). *Rules and representations*. Oxford: Basil Blackwell.
12. Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht/Cinnaminson: Mouton de Gruyter.
13. Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York: Praeger Publishers.
14. Clark, E. V. (2009). *First language acquisition* (2nd ed.). Cambridge/New York: Cambridge University Press.
15. Clark, A., & Lappin, S. (2010). *Linguistic nativism and the poverty of stimulus*. Oxford/Malden, MA: Wiley Blackwell.
16. Cullicover, P. W. (1999). *Syntactic nuts*. Oxford/New York: Oxford University Press.
17. De Marcken, C. G. (1996). *Unsupervised language acquisition*. Ph.D. thesis, MIT.
18. Dominey, P., & Boucher, J. (2005). Learning to talk about events from narrated video in a construction grammar framework. *Artificial Intelligence, 167*(1–2), 31–61.
19. Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah/London: Erlbaum
20. Elman, J. (2001). Connectionism and language acquisition. In *Essential readings in language acquisition*. Oxford: Blackwell.
21. Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology, 31*(1), 41–81.
22. Francis, W., Kučera, H., & Mackie, A. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Harcourt (HMH).
23. Frank, M., Goodman, N., & Tenenbaum, J. (2008). A Bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems, 20*, 457–464.
24. Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition, 13*, 187–222.
25. Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development, 55*, 1535–1540.
26. Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry, 25*, 407–454.
27. Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92* (Vol. 1). New York: IEEE
28. Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*(5), 447–474.
29. Goldberg, A. E. (1999). Emergence of the semantics of argument structure constructions. In *The emergence of language* (Carnegie Mellon Symposia on Cognition Series, pp. 197–212). Mahwah: Lawrence Erlbaum Associates
30. Grünwald, P. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, & G. Scheler (Eds.), *Connectionist, statistical and symbolic approaches to learning for natural language processing* (Lecture Notes in Computer Science, Vol. 1040, pp. 203–216). Berlin/New York: Springer.
31. Hsu, A. S., & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science, 34*(6), 972–1016.
32. Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.
33. Keller, B., & Lutz, R. (1997). Evolving stochastic context-free grammars from examples using a minimum description length principle. In *Workshop on Automata Induction Grammatical Inference and Language Acquisition, ICML-97*. San Francisco: Morgan Kaufmann Publishers
34. Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). *Language Research, 28*(1), 1–13.
35. Legate, J., & Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review, 19*(1/2), 151–162.
36. Leonard, L. (2000). *Children with specific language impairment*. Cambridge: MIT Press.

37. Li, M., & Vitányi, P. M. B. (1995). Computational machine learning in theory and praxis. In J. van Leeuwen (Ed.), *Computer science today* (Lecture notes in computer science, Vol. 1000). Heidelberg: Springer.
38. MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language development: Syntax and semantics* (Vol. 1, pp. 73–136). Hillsdale, NJ: Lawrence Erlbaum.
39. MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erllbaum.
40. MacWhinney, B. (1993). Connections and symbols: Closing the gap. *Cognition, 49*, 291–296.
41. MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
42. MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language, 31*, 883–914.
43. MacWhinney, B., Bird, S., Cieri, C., & Martell, C. (2004). TalkBank: Building an open unified multimodal database of communicative interaction. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon (pp. 525–528). Paris: ELRA
44. Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition, 46*, 53–85.
45. Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). *Overregularization in language acquisition* (Monographs of the society for research in child development, Vol. 57 (4, Serial No. 228)). Chicago: University of Chicago Press
46. Marcus, M., Santorini, B., & Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.
47. Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
48. McClelland, J. L., Rumelhart, D. E., & The PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2). Cambridge, MA: Bradford Books/MIT Press.
49. Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91–117.
50. Parisse, C., & Le Normand, M. T. (2000). Automatic disambiguation of the morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, and Computers, 32*, 468–481.
51. Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37*, 607–642.
52. Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
53. Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua, 92*, 377–410.
54. Pullum, G., & Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review, 19*(1/2), 9–50.
55. Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–471.
56. Roy, D. (2009). New horizons in the study of child language acquisition. In *Proceedings of Interspeech 2009*, Brighton. Grenoble: ISCA
57. Rumelhart, D., & McClelland, J. (1987). Learning the past tenses of English verbs: Implicit rules or parallel distributed processing. *Mechanisms of language acquisition* (pp. 195–248). Hillsdale: Erlbaum.
58. Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition* (pp. 25–32). Prague: Association for Computational Linguistics.
59. Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language, 37*(03), 705–729.
60. Steedman, M., Baldridge, J., Bozsahin, C., Clark, S., Curran, J., & Hockenmaier, J. (2005). Grammar acquisition by child and machine: The combinatory manifesto. *Invited Talk at the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor.

61. Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632.
62. Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318.
63. Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209–253.
64. Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.
65. Tomasello, M., Akhtar, N., Dodson, K., & Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language, 24*(02), 373–387.
66. Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar*. Ph.D. thesis, Computer Laboratory, University of Cambridge.
67. Yang, C. (2002). *Knowledge and learning in natural language*. Oxford/New York: Oxford University Press.
68. Yu, C., & Ballard, D. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing, 70*(13–15), 2149–2165.
69. Yu, C., & Smith, L. (2006). Statistical cross-situational learning to build word-to-world mappings. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, Vancouver. Citeseer.

Part I
Methods and Tools for Investigating Phonetics and Phonology

Phon: A Computational Basis for Phonological Database Building and Model Testing

Yvan Rose, Gregory J. Hedlund, Rod Byrne, Todd Wareham, and Brian MacWhinney

Abstract This paper describes Phon, an open-source software program for the transcription, coding, and analysis of phonetically-transcribed speech corpora. Phon provides support for multimedia data linkage, utterance segmentation, multiple-blind transcription, transcription validation, syllabification, and alignment of target and actual forms. All functions are available through a user-friendly graphical interface. This program provides the basis for the building of PhonBank, a database project that seeks to broaden the scope of CHILDES into phonological development and disorders.

1 Introduction

The topic of this chapter may appear as a slight oddity in the context of the current publication. While most of the contributions to this volume focus on computational methods applied to language learning problems, our paper centers on a recently-introduced tool for the building of phonetically-transcribed speech corpora. This is relevant in a number of respects. Empirical studies of natural language and language

Y. Rose (✉) · G.J. Hedlund
Department of Linguistics, Memorial University of Newfoundland
St. John's, NL A1B 3X9, Canada
e-mail: yrose@mun.ca; ghedlund@mun.ca

R. Byrne · T. Wareham
Department of Computer Science, Memorial University of Newfoundland
St. John's, NL A1B 3X5, Canada
e-mail: rod@mun.ca; harold@mun.ca

B. MacWhinney
Department of Psychology, Carnegie Mellon University
Pittsburgh, PA 15213, USA
e-mail: macw@cmu.edu

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_2, © Springer-Verlag Berlin Heidelberg 2013

acquisition will always be required in most types of linguistic research, as these studies provide the basis for describing languages and linguistic patterns. In addition to providing us with baseline data, corpora also allow us to test models of various kinds, be they theoretical, neurological, psychological or computational. However, the building of natural language corpora is an extremely tedious and resource-consuming process, despite tremendous advances in data recording, storage, and coding methods in recent decades.

Thanks to corpora and tools such as those developed in the context of the CHILDES project (http://childes.psy.cmu.edu/), computational scientists interested in morphology and syntax have enjoyed convenient and powerful methods for analysing the morphosyntactic properties of natural languages and their acquisition by first and second language learners. In the area of phonetics, the Praat system (http://www.fon.hum.uva.nl/praat/) has expanded our abilities to test optimality-theoretic as well as neural network-based learning models, in addition to providing a breadth of support in the areas of speech analysis and synthesis.

In this rapidly-expanding software universe, phonologists interested in the organisation of sound systems (e.g. phones, syllables, stress and intonational patterns) and their acquisition had not enjoyed the same level of computational support prior to the inception of the PhonBank project within CHILDES. There was no developed platform for phonological analysis and no system for data sharing. This situation negatively affected the study of natural language phonology and phonological development. It also undermined potential studies pertaining to interfaces between various components of the grammar or the elaboration of computational models of language and language development.

It is widely accepted that a spoken utterance consists of more or less one sentence. Utterances can contain one or more phonological phrases, which can serve as reference domains for intonational purposes or relate to independent aspects of syntactic constituency. Phonological phrases are typically made of series of one or more prosodic words and associated morphemes, with each of these meaningful units consisting of syllables which can themselves be broken down into individual phones. This general grammatical organisation allows us to make reference to factors that link the smallest phonological units to morphological and/or syntactic levels of grammatical patterning. For example, in English, the phonological phrase, a domain that constrains phonological phenomena such as intonation, must typically be described using syntactic criteria; in a similar way, the analysis of stress patterns in this language requires references to large-domain morphological boundaries (e.g. [22]). Studying the acquisition of these grammatical structures, and of their phonological components, can help us understand how linguistic knowledge emerges in the grammar of a language learner.

In this paper we discuss Phon, an innovative open-source software program that offers significant methodological advances in research in phonology and phonological development. On the one hand, Phon provides a powerful and flexible solution for phonological corpus building and analysis. On the other hand, its

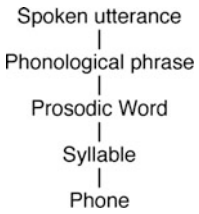

Fig. 1 General organisation of a spoken utterance

ability to integrate with other open-source software facilitates the construction of complete analyses across all levels of grammatical organisation represented in Fig. 1. Although the primary target group for this tool was originally L1 researchers, the core functions of Phon are equally valuable to other speech researchers who are interested in analysing language variation of any type (e.g. cross-dialectal variation, L2 speech, speech pathologies, evolution of consonant inventories; on this last application, see Mukherjee et al. in this volume).

The paper is organised as follows. In Sect. 2, we discuss the general motivation behind the Phon project. In Sect. 3, we describe the functionality supported in Phon. In Sect. 4, we focus on the query and reporting systems that are built into the application. We then summarize in Sect. 5 currently planned extensions to Phon, including both the integration of acoustic data analyses and a greatly expanded database query functionality that will ultimately assist in both language acquisition model testing and derivation. Concluding remarks are offered in Sect. 6.

2 The PhonBank Project

PhonBank, one of the latest initiatives within the CHILDES project, focuses on the construction of corpora suitable for phonological and phonetic analysis. In this section we first describe the goals of PhonBank. We then describe Phon, the software program designed to facilitate this endeavor.

2.1 PhonBank

The PhonBank project seeks to broaden the scope of the current CHILDES system to include the analysis of phonological development in first and subsequent languages for learners with and without language disorders. To achieve this goal, we have created a new phonological database called PhonBank and a program called Phon to facilitate the analysis of PhonBank data. Using these tools, researchers are in a position to conduct a series of developmental, crosslinguistic, and methodological analyses based on large-scale corpora.

2.2 *Phon*

Phon consists of inter-connected modules that offer functionality to assist the researcher in important tasks related to corpus transcription, coding and analysis. (The main functions supported are discussed in the next section.)

The application is developed in Java and is packaged to run on Mac OS X and Windows platforms that support Java 1.6.[1] Phon is Unicode-compliant, a required feature for the sharing of data transcribed with phonetic symbols across computer platforms. Phon can share data with programs which utilize the TalkBank XML schema for their documents such as those provided by the TalkBank and CHILDES projects.

Phon was introduced approximately 5 years ago (see [19]). Since then, we have thoroughly revised significant portions of the code to refine the functionality, ensure further compatibility with other TalkBank-compliant applications, and streamline the interface for better user experience and improved support for the general workflow involved in phonological corpus building. We also added novel and innovative functionality for corpus query and reporting. An advanced beta version of this application is publicly available online as a free download directly from the CHILDES website (http://childes.psy.cmu.edu/).

3 Phon

The general interface of Phon is exemplified in Fig. 2. It consists of a series of view panels, each of which supports particular aspects of corpus manipulation (e.g. session-level information, orthographically or phonetically transcribed data, other data annotations). In Fig. 2, three view panels are displayed (Record Data, Syllabification and Alignment, and the Waveform of the speech segment transcribed in this record). Additional view panels can be docked horizontally or vertically, or superposed as tabbed interfaces within single docks. This user-configurable interface is one of the key improvements brought to the current version. Using this interface, the user can perform a series of tasks related to the building of phonological corpora:

- Media linkage and segmentation.
- Data transcription and validation (including support for multiple-blind transcriptions).
- Segmentation of transcribed utterances (into e.g. phrases, words).
- Labeling of transcribed forms for syllabification.

[1] Support for the Unix/Linux platform is currently compromised, primarily because of licensing issues related to the multimedia functions of the application.

Fig. 2 Phon general interface

- Phone and syllable alignment between target (expected) and actual (produced) forms.

In the next subsections, we describe the main functions supported by the application.

3.1 Project Management

The building of a corpus of transcribed phonological data often requires the combination of a number of data transcripts, be they from a single learner over a period of time or from multiple learners. Phon offers functions to create and manage sets of data transcripts, following a general corpus structure whereby a project contains one or many corpora, each of which contains a set of data transcripts. Session transcripts can be copied or moved across corpora or project files through the Project Manager interface.

3.2 Media Linkage and Segmentation

Linkage of multimedia data and subsequent identification of the portions of the recorded media that are relevant for analysis are now available directly from the application's main interface. These tasks follow the same logic as similar systems in programs like CLAN (http://childes.psy.cmu.edu/clan/). A transcript in Phon generally corresponds to a data recording session. The created media segments can be played back directly from the graphical user interface (GUI). Whenever needed, the user can also fine-tune the segments start and/or end time values, a task made much easier with the incorporation of waveform visualisation.

3.3 Data Transcription

As we saw in Fig. 2, the Session Editor incorporates in a single interface access to data transcription and annotation, transcription segmentation, syllabification and alignment. Phon also provides support for an unlimited number of user-defined fields that can be used for various kinds of textual annotations that may be relevant to the coding of a particular dataset. All data tiers can be ordered to accommodate specific data visualisation needs. Support for tier-specific fonts is also provided, a feature particularly useful for work based on non-Roman data transcripts. Phonetic transcriptions are based on the phonetic symbols and conventions of the International Phonetic Association (IPA). A useful IPA character chart is easily accessible from within the application, in the shape of a floating window within which IPA symbols and diacritics are organised into intuitive categories. This chart facilitates access to the IPA symbols for which there is no keyboard equivalent.

Target and actual IPA transcriptions are stored internally as strings of phonetic symbols. Each symbol is automatically associated with a set of descriptive features generally accepted in the fields of phonetics and phonology (e.g. bilabial, alveolar, voiced, voiceless, aspirated) [13]. These features are extremely useful in the sense that they provide series of descriptive labels to each transcribed symbol. The availability of these labels is essential for research involving the grouping of various sounds into natural classes (e.g. voiced consonants; non-high front vowels). The built-in set of features can also be reconfigured to fit particular research needs through the query system (see Sect. 4.5 for further discussion).

Phon is also equipped with functionality to automatically insert IPA Target transcriptions based on the orthographic transcriptions. Citation form IPA transcriptions of these words are currently available for Catalan, German, English, French, Icelandic, Italian, Dutch and Spanish. IPA dictionary files from these various languages were generously provided by independent open-source projects and subsequently formatted for use into Phon.

In cases when more two or more pronunciations are available from the built-in dictionaries for a given written form (e.g. the present and past tense versions

of the English word ‘read’), the application provides a quick way to select the required form. Of course, idealized citation forms do not provide accurate fine-grained characterisations of variations in the target language (e.g. dialect-specific pronunciation variants; phonetic details such as degree of aspiration in obstruent stops). They, however, typically provide a useful general baseline against which patterns can be identified.

3.4 Multiple-Blind Transcription and Transcript Validation

Phon offers a fully-integrated system for multiple-blind, consensus-based IPA transcriptions. Multiple-blind transcription is in essence identical to the double-blind protocol: it consists of the IPA transcription of recorded utterances by two or more (hence, multiple) transcribers. Within Phon, the IPA transcribers are effectively ‘blinded’ from each other’s transcriptions in that they must perform their IPA transcriptions without being able to visualise the transcriptions of other transcribers. Each IPA transcriber logs into the blind transcription interface using a specific username. Upon login, a transcriber can visualise the regular corpus data records, including orthographic transcriptions and other annotations, with the crucial exception that the visible IPA transcription tiers are unique to each transcriber.

After the blind transcriptions are performed, they are then ready for the next step in the workflow, which consists of consensus-based transcript validation. This step is necessary as, under the blind transcription protocol, none of the user-specific transcriptions can immediately be considered valid for research. We developed an interface within Phon which facilitates record-by-record comparisons of the blind transcriptions. Using this interface, a team of two (or more) transcript validators can listen to the record’s speech segment, and then visualise in parallel all corresponding transcriptions produced by the blind transcribers. The transcription deemed the most accurate by consensus between the transcript validators is then selected with a simple mouse click. Whenever necessary, the selected transcription can be further adjusted according to the details noticed by the transcript validators during the comparison process. While this method is relatively onerous both in time and human resources, its combined steps (blind transcription followed by consensus validation) help to maximise transcription reliability for research purposes.

Employing blind transcription and its associated validation systems is optional. If the user decides not to perform blind transcriptions, the phonetic transcriptions are entered directly into the transcript and, as such, do not require subsequent validation. Similarly, the decision to protect each set of blind transcriptions with a password, which may be overkill in many situations, is left to the user. Note as well that only data which have been validated or directly entered into the transcript can be further annotated or compiled. Non-validated blind transcriptions are saved as part of the project file but cannot be used for research. Whichever the mode of entry into the session editor (multi-blind or direct), the interface for data entry remains identical.

Of course, as with anything related to phonetic transcription, whichever method selected by the user is no panacea. Regardless of the amount of care put into it, and in spite of its crucial role in creating readable transcripts of spoken forms, the symbolic representation of speech sounds remains a methodological compromise. Nonetheless, at all steps involving the transcription of spoken utterances into IPA notation and/or the validation of phonetic transcripts, the user (transcriber or validator) can always export the relevant speech samples as individual sound clips for visualisation in speech analysis software programs for further assessment of the properties of the speech signal. This functionality further contributes to the attainment of the most representative data transcripts possible.

3.5 Transcribed Utterance Segmentation

Researchers often wish to divide transcribed utterances into specific domains such as the phrase or the word. Phon provides basic functionality to address this need by incorporating a text segmentation module that enables the identification of strings of symbols corresponding to such morphosyntactic and phonological domains, which we loosely call ‘word groups’. An important feature of word grouping is that, if used, it strictly enforces a logical organisation between Orthography, IPA Target and IPA Actual tiers, the latter two being treated as daughter nodes directly related to their corresponding parent bracketed form in the Orthography tier. Word groups are also supported in user-defined tiers. This system of tier dependency offers several analytical advantages, for example for the identification of patterns that can relate to a particular grammatical category or position within the utterance.

3.6 Syllabification Algorithm

Once IPA transcriptions are entered into the transcript, Phon performs syllable-level annotation automatically: segments are assigned descriptive syllable labels (visually represented with colors) such as ‘onset’ or ‘coda’ for consonants and ‘nucleus’ for vowels, as can be seen in Fig. 3.

Our general approach to syllable-level annotations is based on models of syllable representation developed within generative phonology, which provide a particularly useful framework for its focus on structural description (e.g. [10, 21]; see [5] and [6] for applications to child phonology). However, because some degree of controversy exists in both phonetic and phonological theory regarding the very notion of syllabification and the types of syllable constituents allowed across formal models, the algorithm can be easily parameterized to suit various models. As can be seen in Fig. 3, descriptive models supported by Phon can be highly articulated, with positions such as initial (left) and final (right) appendices, or less refined, for example with all pre-vocalic consonants as onsets and post-vocalic ones as codas.

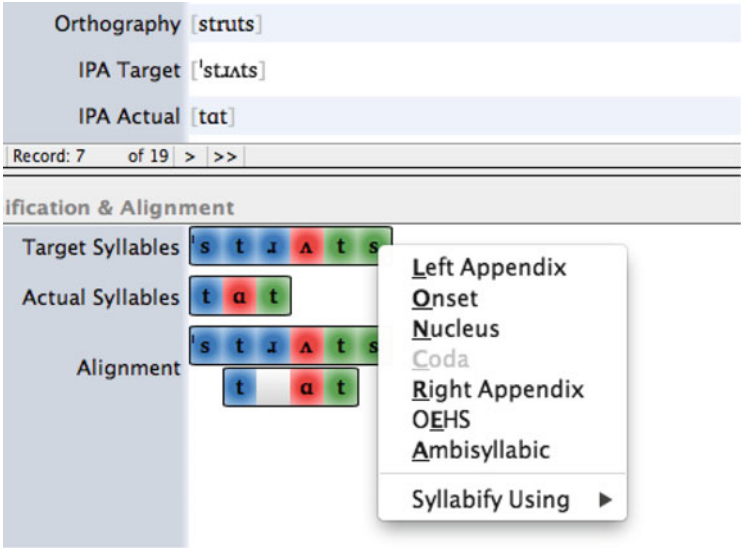

Fig. 3 Syllable-level annotations

Note as well that the use of syllable-level annotation can be used in a number of ways. On the one hand, the availability of different parameter settings makes it possible to test various hypotheses for any given dataset, for example, about formal distinctions concerning the status of on-glides in English (e.g. [3]). On the other hand, labels can be used in a strictly descriptive fashion, to ease tasks such as precise data compilation, but yet have no formal implications for the researcher's theoretical approach (and related analysis).

Syllabification algorithms are provided for several languages, with multiple algorithms readily available for English, French and Dutch. Additional algorithms (for other languages or based on different assumptions about syllabification) can easily be added to the program, upon request. These algorithms use a scheme based on a composition-cascade of seven deterministic FSTs (Finite State Transducers). This cascade takes as input a sequence of phones and produces a sequence of phones and associated syllable-constituent symbols, which is subsequently parsed to create the full multi-level prosodic annotation. The initial FST in the cascade places syllable nuclei. Subsequent FSTs establish and adjust the boundaries of associated syllable onset and offset domains. Changes in the definition of syllable nuclei in the initial FST and/or the ordering and makeup of the subsequent FSTs give language-specific syllabification algorithms. To ease the development of this cascade, initial FST prototypes were written and tested using the Xerox Finite-State Tool (xFST) [1]. However, following the requirements of easy algorithm execution within and integration into Phon, these FSTs were subsequently coded in Java. To date, the implemented algorithm has been tested on corpora from English and French, and has obtained extremely high accuracy rates.

Occasionally, the algorithm may produce spurious results or flag symbols as unsyllabified. This is particularly true in the case of IPA Actual forms produced by young language learners, which sometimes contain strings of sounds that are not attested in natural languages. Since syllabification annotations are generated on the fly upon transcription entry within the IPA Target or IPA Actual tiers, the researcher can quickly verify all results and modify them through a contextual menu (represented in Fig. 3) whenever needed. Segments that are left unsyllabified are available for all queries on segmental features and strings of segments, but are not available for queries referring to aspects of syllabification.

The syllabification labels can then be used in database query (for example, to access specific information about syllable onsets or codas). In addition, because the algorithm is sensitive to main and secondary stress marks and domain edges (i.e. first and final syllables), each syllable identified is given a prosodic status and position index. Using the search functions, the researcher can thus use search criteria as precisely defined as, for example, complex onsets realised in word-medial, secondary-stressed syllables. This level of functionality is central to the study of several phenomena in phonological acquisition that are determined by the status of the syllable as stressed or unstressed, or by the position of the syllable within the word (e.g. [9]).

3.7 Alignment Algorithm

After syllabification, a second algorithm performs automatic, segment-by-segment and syllable-by-syllable alignment of IPA-transcribed target and actual forms. Building on featural similarities and differences between the segments in each syllable and on syllable properties such as stress, this algorithm automatically aligns corresponding segments and syllables in target and actual forms. It provides alignments for both corresponding sounds and syllables. For example, in the target-actual word pair ‘apricot’ > ‘apico’, the algorithm aligns the phones contained in each corresponding syllable, as illustrated in Fig. 4.

In this alignment algorithm, forms are viewed as sequences of phones and syllable-boundary markers. Alignment is performed on the phones in a way that preserves the integrity of syllable-level annotations. This algorithm is a variant of the standard dynamic programming algorithm for pairwise global sequence alignment (see [20] and references therein); as such, it is similar to but extends the phone-alignment algorithm described in [12].

At the core of the Phon alignment algorithm is a function $sim(x, y)$ that assesses the degree of similarity of a symbol x from the first given sequence and a symbol y from the second given sequence. In our $sim()$ function, the similarity value of phones x and y is a function of a basic score (which is the number of phonetic features shared by x and y) and the associated values of various applicable reward

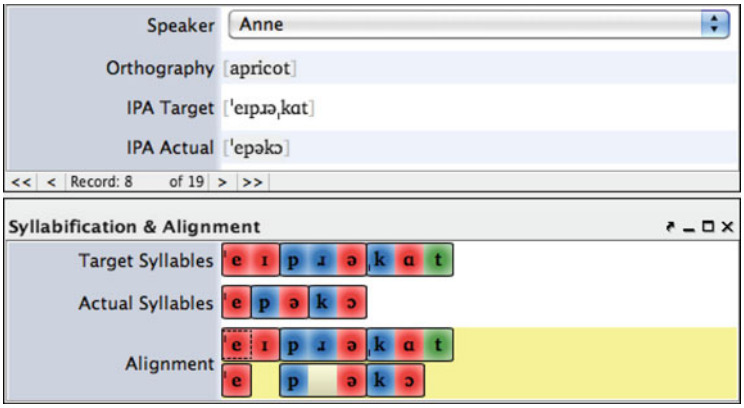

Fig. 4 Phone alignment

and penalty conditions, each of which encodes a linguistically-motivated constraint on the form of the alignment. There are nine such reward and penalty conditions, and the interaction of these rewards and penalties on phone matchings effectively simulates syllable integrity and matching constraints. Subsequent to this phone alignment, a series of rules is invoked to reintroduce the actual and target form syllable boundaries.

A full description of the alignment algorithm is given in [8, 14]. As it is the case with all of the other algorithms for automatic annotation included in the program, the user is able to perform manual adjustments of the computer-generated syllable alignments whenever necessary. This process was made as easy as possible: it consists of clicking on the segment that needs to be realigned and moving it leftward or rightward using keyboard arrows.

The alignment algorithm, as well as the data processing steps that precede it (especially, syllabification), are essential to any acquisition study that requires pair-wise comparisons between target and actual forms, from both segmental and syllabic perspectives. Implicit to the description of the implementation of the syllabification and alignment functions is a careful approach whereby specialized algorithms are implemented in ways that facilitate data annotation; because every result generated by the algorithms can be modified by the user, no ensuing compilation of these data directly depends upon them. The user thus has complete control on the processing of the data being readied for analysis. After extensive testing on additional types of data sets, we will be able to optimize their degree of reliability and then determine how they can be used in truly automated analyses.

Once the data are processed through the modules described in the preceding subsections, they are ready to be used for research. We describe in the next section the search and reporting functions supported by Phon.

4 Database Query

The newly-introduced query system can be loosely described as a plug-in system based on JavaScript. Built-in query scripts are provided for general data query purposes. Using these scripts, the researcher can identify records that contain:

- Phones and phone sequences (defined with IPA symbols or descriptive feature sets).
- Syllable types (e.g. CV, CVC, CGV, etc. where C = consonant, V = vowel, and G = glide).
- Word types (e.g. number of syllables; stress patterns).
- IPA Target-Actual phone and syllable-level comparisons, obtained through phone and syllable alignment (e.g. phone substitutions; complex onsets reduction; syllable epenthesis).

Beyond these built-in search functions, the user can create search scripts without any need to reprogram the application. In this section, we discuss the processes of executing a query using this system, reviewing the results, and creating a report of (or an export based on) those results. This discussion also briefly covers how queries are specified using JavaScript and how to find more information about this system.

4.1 Terminology

Several terms used in the following sub-sections must first be defined

- **Query:** The set of criteria (or pattern) used to match results. Each query executed in Phon is given a unique ID.
- **Search:** The execution of a query on a particular session. Each search is also given a unique ID.
- **Search Metadata:** A set of key/value pairs which contains data particular to the search being performed.
- **Result:** A result is a single instance of the given patten found in a session.
- **Result Metadata:** A set of key/value pairs which contains data related to a result.
- **Query History:** A history of all queries performed for the current project. Queries can be 'starred' for later reference and be re-opened at later dates. The Query History window can be accessed via the Project >> Query History menu option in the Project Manager window.
- **Script Editor:** Phon provides a basic script editor for creating custom queries. The Script Editor can be accessed via the Project >> Script Editor menu option in the Project Manager window.

4.2 Executing a Query

Phon's search and report functions are separated into three phases:

1. Specifying Query: Query parameters are entered and the search is executed on one or more selected sessions.
2. Review Intermediate Results: All results are stored in a relational database. Result sets can be opened in Phon and modified using the Session Editor.
3. Generate Report: Results can be exported as reports into a variety of formats for use in other applications and further analysis.

Phase 1: Specifying Query Phon provides several stock searches with the installer. Each of these searches has a form which can be invoked by using the Project >> Search menu (for project-level searching) and View >> Search menu (within the Session Editor). The stock searches included with Phon are:

- **Aligned Groups:** A search for tiers which are organised into phonetic groups. This is useful for performing searches where special coding is used inside user-defined tiers which is associated with data in the Orthography or IPA tiers.
- **Aligned Phones:** This search is provided for searching patterns in the alignment data of records. Patterns are specified using Phon's phonex language for matching phone sequences.
- **CV Sequence:** Used for searching CV(G) patterns in IPA data.
- **Data Tiers:** This search is provides for generic searching of any tier.
- **Harmony:** A special function search for locating instances of harmony (consonant and/or vowel) in aligned IPA forms.
- **Metathesis:** A special function search for locating instance of metathesis in aligned IPA forms.
- **Word Shapes:** Used for searching stress patterns in words.

Each search form includes options for selecting group, word, and syllable position; syllable stress; and participant name and age, where applicable. Once the options in the form have been specified, the query can be executed on one or more sessions in the currently open project (or on the current session if initiated from the Session Editor.)

Phase 2: Review Intermediate Results Once all searches have been completed result sets can be reviewed using the Session Editor. Results are displayed in a table with corresponding metadata. Results can be removed from the result set; however such removals cannot be undone. This process is especially useful for queries which can potentially return false-positives, such as the Harmony and Metathesis queries.

Phase 3: Generate Report Query reports can come in two formats: a flat-export Comma-Separated Values (CSV) file; and a printable format which can be exported into a variety of formats. The user can choose to create a query report once project-level searches have been completed. Access to the reporting function is also available in the search list of the Query History window.

CSV reports can organise all results in either a single file or a set of files (each of which corresponds to an individual session). The user also has the option of selecting the columns they want included in the report. Columns for session data, speaker information, tier data and result data are available. This report type is most useful for inputting data into other applications (such as SPSS) for analysis.

For more detailed reporting Phon provides a configurable report template which can generate printable reports in PDF, Microsoft Word/Excel, OpenOffice, and formatted CSV. When creating these reports a default option is provided for the user. This default option can be changed by adding/removing report sections in the provided interface. Currently there are sections for printing parameters used for a query, search summary, comments, inventories, and result listings. Each section has configuration options allowing further customization of the report.

4.3 Creating a Query

Phon queries are written in a language called JavaScript. Advanced users can take advantage of having a full programming language for creating customized queries. Essentially any query can be defined using this system but there are restrictions as to what can constitute a result. The application programming interface (API) for queries can be found by using the help button in the Script Editor.

A minimal script for a Phon query implements the function `query_record(record)`. This method is executed once for each record in a session. The provided variable 'record' is a reference to the current record. A global variable 'results' is also provided for adding results to the current search's result set. Optionally, the user can implement the functions `begin_search(session)` and `end_search(session)` which are executed at the beginning and end of a search, respectively. These methods are useful for initializing and reporting any custom global variables.

4.4 An Illustrative Example

In this section we present a concrete illustration of data representation and querying within Phon. This illustration draws on the Goad-Rose corpus of Quebec French development available through PhonBank, aspects of which are analysed in Rose [18]. First, we illustrate in Fig. 5 an early production of the name "Gaspard".

As we can see in this illustration, Phon enables the identification of segmental discrepancies between the target (model) pronunciation and its actual production by the French learner through an alignment of all IPA Target phones with their IPA Actual counterparts, thereby setting the stage for investigations of the segmental and

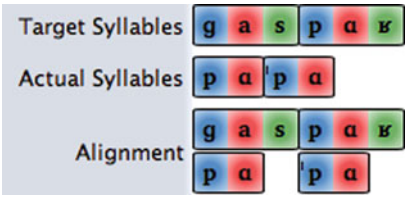

Fig. 5 Pronunciation of French name "Gaspard"

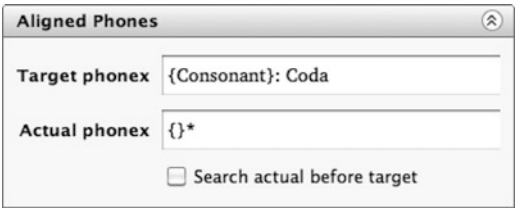

Fig. 6 Aligned Phones query: realisation of target codas

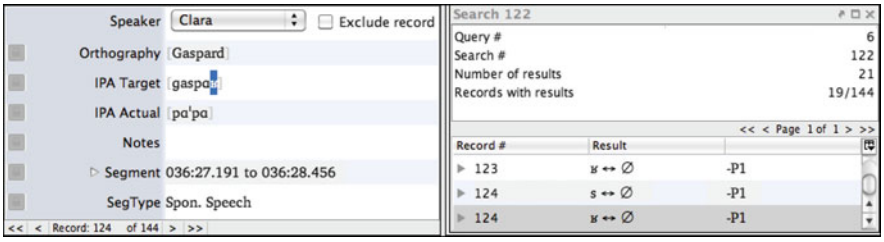

Fig. 7 Search results visualised in the transcript editor

syllabic properties of the child production (e.g. target [g] produced as [p]; deletion of both syllable-final consonants). For example, focusing on patterns of coda (syllable-final) consonant deletion, one can use the Aligned Phones search system to identify all of the relevant cases in the database, as shown in Fig 6.

In this example, the user invokes the phonex language to search for target consonants in syllable codas and associated productions (or deletions) in the child's forms. As we can see in Fig. 7, the application identifies matching patterns, each of which can be visualised from the application's GUI.

As mentioned in Phase 3 of Sect. 4.2 above, such results can be compiled for entire recording sessions or corpora (collections of recording sessions) in the form of reports generated in various formats. Portions of Phon-generated reports are illustrated in the next two figures from a report generated in the Excel format. All such reports are also divided into specific sections. For example, as illustrated in Fig. 8, sections reporting general inventories of results are useful to observe general trends in the data. In this particular case, we can see that all continuant consonants

Fig. 8 Phon-generated data report: inventory of results

Inventory

Result format	IPA Target ↔ IPA Actual
Result	*Count*
ʁ ↔ ∅	10
m ↔ m	1
j ↔ ∅	6
s ↔ ∅	4

Fig. 9 Phon-generated data report: record-by-record result listing

Result Listing

Record	Result format	Result
31	IPA Target ↔ IPA Actual	ʁ ↔ ∅
	Tier data:	
	Orthography	[ourson]
	IPA Target	[ʊʁsɔ̃]
	IPA Actual	[ʊˈsɔ̃]

in coda undergo deletion (e.g. all ten instances of target rhotics in coda), while coda [m], a stop consonant, is realised in a target-like fashion.

In order to fully appreciate the extent of such patterns, the user can also study each instance of a given pattern through individual result listings such as the one illustrated in Fig. 9. While this example only contains data from three tiers (Orthography, IPA Target and IPA Actual), all tiers contained in data records can be listed in the reports.

While the above examples provide a simple illustration of the query and reporting system implemented in Phon, many more types of query and data reports can be specified by the user. Through combinations of quantitative information (as in Fig. 8) and associated qualitative characterizations (as in Fig. 9), the user can achieve the desired level of observational detail to address various types of research hypotheses.

4.5 Additional Information

More information on searching in Phon can be found in the application manual available via the Help menu. A support forum is also available, which can be accessed at http://phon.ling.mun.ca/phontrac/discussion/3. For additional query scripts visit http://phon.ling.mun.ca/phontrac/wiki/search/scriptlibrary. Finally, as mentioned in Sect. 3, support is also provided for data compilations based on particular feature sets for transcribed phones. Information on this topic can be found at http://phon.ling.mun.ca/phontrac/wiki/search/customfeatures.

5 Future Projects

As described above, Phon provides all the functionality required for corpus building as well as a versatile system for data extraction. In future versions, we plan to incorporate an interface for the management of acoustic data and fuller support for data querying and searching; the latter can, among other things, be used to create systems for testing and deriving language acquisition models. We will discuss all of these plans in the following subsections.

5.1 Interface for Acoustic Data

In order to facilitate research that requires acoustic measurements, Phon will interface fully with Praat [2], a software program designed for acoustic analysis of speech sounds. Using conduits between Praat and Phon, researchers that use these programs will be able to take advantage of some of Phon's unique functions and, similarly, researchers using Phon will be able to integrate acoustic measurements for both corpus preparation and data analysis.

We will first develop an interface within Phon for the alignment of phonetic transcriptions with their corresponding waveforms and spectrograms. This process will be semi-automated using the CSLU Toolkit (http://cslu.cse.ogi.edu/toolkit), which provides a method for aligning phonetic transcriptions with their corresponding spectrographic representations. Researchers will also be able to use similar Praat-compatible plug-ins such as EasyAlign, which can be accessed through http://latlcui.unige.ch/phonetique/. The transcription-spectrogram alignment will provide the start and end points of data measurements for each sound or sound sequence targeted by the researcher. The researcher can then activate a command to send the relevant portion of the recorded media for analysis in Praat, which can compute a wide variety of acoustic analyses, such as F0 and formant tracking and spectral analysis through FFT or LPC. After acoustic analysis, Phon will import the results into an interface that will accommodate acoustic measurement data.

The system described above will offer unprecedented support for investigations requiring the combination of refined phonological classifications and detailed acoustic characterizations of developmental data. Among other advantages, this system will offer a means to systematically verify phonetic transcriptions of recorded speech, through mapping impressionistic transcriptions with their acoustic correlates. It will also enable systematic extractions and compilations of acoustic measurements of speech sounds relative to their positions within the spoken utterance. For example, the researcher will be able to study the development of vocalic systems by simultaneously compiling longitudinal acoustic data relative to prosodic positions such as stressed versus unstressed syllables. As a result, researchers that utilize Praat will be able to take advantage of some of Phon's unique functions and, similarly, researchers using Phon will be able to take advantage of the functionality of both Praat and Phon.

5.2 Extensions of Database Query Functionality

The search and report functions described in Sect. 4.2 provide simple and flexible tools to generate general assessments of the corpus or detect and extract particular phonological patterns occurring in specified data tiers in individual sessions. However, to take full advantage of all of the research potential that Phon offers, a more powerful query system is required. The first steps towards such a system will involve modifying the existing query language to include (1) standard statistical functions and (2) methods for specifying queries that incorporate the acoustic data described in the previous subsection. This will enable precise initial assessments of developmental data within and across corpora of language learners or learning situations.

More comprehensive assessments are possible if the query mechanism is further modified to allow the specification and matching of richer types of patterns. One such type of particular interest is a pattern that describes a (possibly summarized) portion of the time-series of sessions comprising the longitudinal data for a particular language learner. Each such pattern is effectively a hypothesis about the nature and course of language acquisition. The hypotheses associated with these patterns treat linguistic phenomena of variable extent, from local features of language acquisition that occur in a particular time-interval, e.g., the so-called 'vocabulary spurt', to global characterizations of the whole acquisition process, e.g., the acquisition order of the target phone inventory.

Given such a pattern and a learner time-series, the degree to which the pattern matches the time-series corresponds to the degree with which the hypothesis encoded by that pattern is consistent with and hence supported by that time-series. Alternatively, two learner time-series could be matched against each other to assess their similarity and hence the degree to which those learners are following the same developmental path. Such matches can be done with variants of the target-actual form alignment function described in Sect. 3.7. Many types of time-series pattern matching have been defined and implemented within computational molecular biology (see [7, 20] and references) and temporal data mining (see [11, 15, 17] and references); it seems likely that some of these can be either used directly or modified for the purposes of language acquisition research.[2]

[2] Previous experience in computational molecular biology and data mining suggests that, given the large amounts of data involved, various specialized algorithmic techniques will probably have to be invoked to allow time-series pattern matching to run in practical amounts of time and computer memory. The typical approach described in [11] is to simplify the given data, derive approximate analysis-results relative to this simplified data, and (hopefully with minimal effort) reconstruct exact analysis-results relative to the original data. However, there may be other options, such as using so-called fixed-parameter tractable algorithms [4, 16] whose running times are impractical in general but efficient under the restrictions present in learner time-series datasets.

While useful in itself, such a time-series pattern matching capability is the building block of even more exotic analyses, e.g.,

- **Language Acquisition Model Testing:** Given appropriate formalizations of language acquisition models as algorithms that produce 'actual' output analogous to that produced by learners, such models can be automatically evaluated against learner time-series stored in Phon (in a manner analogous to the 'Learn' function in Praat) using functions such as:
  - Run an arbitrary language learning algorithm.
  - Compare the results of the grammar produced by such a language learning algorithm against actual language data.
  - In the event that the learning algorithm provides a sequence of grammars corresponding to the stages of human language learning, compare the results of this sequence of grammars against actual longitudinal language data.

  By virtue of its software architecture, form-comparison routines, and stored data, Phon provides an excellent platform for implementing such an application. Running arbitrary language learning algorithms can be facilitated using a Java API/interface-class combination specifying subroutines provided by Phon, and the outputs of a given model could be compared against target productions stored in Phon using either the alignment algorithm described in Sect. 3.7 or the more general time-series pattern matching algorithms described above.
- **Language Acquisition Model Derivation:** Consider applying time-series pattern matching as described above in reverse – namely, given a set of two or more learner time-series, find those patterns that are best supported by and hence characteristic of the those time-series. Such patterns may be used as developmental benchmarks for deriving language acquisition models or (in the case of very rich types of time-series patterns) function as models themselves.

Analyses such as those sketched above would be much more comprehensive than what has been the norm thus far in the field, especially given past problems encountered in verifying let alone deriving trustworthy models of language acquisition relative to small (and possibly wildly unrepresentative) sets of learners. However, given the diversity of analysis techniques available within computational molecular biology and data mining in general, providing a platform like Phon for implementing such analyses may have the (perhaps ultimately more important) long-term effect of introducing previously unimagined analytical possibilities and related research opportunities.

6 Discussion

Phon offers a sound computational foundation for the management of corpus-based research on phonology and phonological development, media linkage and segmentation into transcript-annotated time intervals, multiple-blind IPA transcription,

IPA transcription validation, target (adult) IPA form insertion, automatic phone alignment between target (model) and actual (produced) forms, automatic syllabification, utterance segmentation into smaller units, database query, data import, and data export. Finally, it provides a strong computational foundation for the implementation of additional functions.

Beyond acoustic data analysis capabilities, the order in which new functionalities will be implemented in future versions of Phon is still unclear. For example, the model-testing tool sketched in Sect. 5.2 is ambitious and perhaps premature in some respects, e.g., should we expect the current (or even next) generation of language learning algorithms to mimic the longitudinal behavior of actual language learners? Such issues are especially relevant, given that some language behaviors observed in learners can be driven by articulatory or perceptual factors, the consideration of which implies relatively more complex models. That being said, the above suggests how Phon, by virtue of its longitudinal data, output-form comparison routines, and software architecture, may provide an excellent platform for implementing the next generation of computational language analysis tools.

Acknowledgements We would like thank the co-organisers of the original ACL workshop (namely, Afra Alishahi, Thierry Poibeau, Anna Korhonen and Aline Villavicencio) for their help and support through all the steps that brought us to this publication and Carla Peddle for assistance in preparing the final version presented here. We are also grateful to two anonymous reviewers for their useful feedback. Current development of Phon and PhonBank is supported by the National Institute of Health. Earlier development of Phon was funded by grants from National Science Foundation, Canada Fund for Innovation, Social Sciences and Humanities Research Council of Canada, Petro-Canada Fund for Young Innovators, and the Office of the Vice-President (Research) and the Faculty of Arts at Memorial University of Newfoundland. TW would also like to acknowledge support provided through NSERC Discovery Grant 228104.

References

1. Beesley, K. R., & Karttunen, L. (2003). *Finite-state morphology*. Stanford, CA: CSLI Publications.
2. Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer [Computer program]. Version 5.2.18, retrieved 10 March 2011 from http://www.praat.org/.
3. Davis, S., & Hammond, M. (1995). On the status of onglides in American English. *Phonology, 12*, 159–182.
4. Downey, R. G., & Fellows, M. R. (1999). *Parameterized complexity*. New York: Springer.
5. Fikkert, P. (1994). *On the acquisition of prosodic structure*. Dordrecht: ICG Printing.
6. Goad, H., & Rose, Y. (2004). Input elaboration, head faithfulness and evidence for representation in the acquisition of left-edge clusters in West Germanic. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints on phonological acquisition* (pp. 109–157). Cambridge/New York: Cambridge University Press.
7. Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge/New York: Cambridge University Press.
8. Hedlund, G. J., Maddocks, K., Rose, Y., & Wareham, T. (2005). Natural language syllable alignment: From conception to implementation. In *Proceedings of the Fifteenth Annual Newfoundland Electrical and Computer Engineering Conference (NECEC 2005)* http://www.ucs.mun.ca/~yrose/Research/Publications/files/2005-HedlundEtAl-SyllAlign.pdf.

9. Inkelas, S., & Rose, Y. (2007). Positional neutralization: A case study from child language. *Language, 83*, 707–736.
10. Kaye, J., & Lowenstamm, J. (1984). De la syllabicité. In *Forme sonore du langage* (pp. 123–161). Paris: Hermann.
11. Keogh, E. (2008). Indexing and mining time series data. In: S. Shekhar & H. Xiong (Eds.), *Encyclopedia of GIS* (pp. 493–497). New York: Springer.
12. Kondrak, G. (2003). Phonetic alignment and similarity. *Computers in the Humanities, 37*, 273–291.
13. Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Cambridge, MA: Blackwell.
14. Maddocks, K. (2005). *An effective algorithm for the alignment of target and actual syllables for the study of language acquisition*. B.Sc.h. thesis. Memorial University of Newfoundland.
15. Mitsa, T. (2010). *Temporal data mining*. Boca Raton, FL: Chapman and Hall/CRC.
16. Niedermeier, R. (2006). *Invitation to fixed-parameter algorithms*. Cambridge/New York: Oxford University Press.
17. Roddick, J. F., & Spiliopoulou, M. (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering, 14*, 750–767.
18. Rose, Y. (2000). *Headedness and prosodic licensing in the L1 acquisition of phonology*. Ph.D. dissertation. McGill University.
19. Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G. J., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing phon: A software solution for the study of phonological Acquisition. In *Proceedings of the 30th Boston University Conference on Language Development* (pp. 489–500). Somerville, MA: Cascadilla Press.
20. Sankoff, D., & Kruskal, J. B. (Eds.). (1983). *Time warps, string edits, and macromolecules: The theory and practice of string comparison*. Reading, MA: Addison-Wesley.
21. Selkirk, E. (1982). The syllable. In *The structure of phonological representation* (pp 337–385). Dordrecht: Foris.
22. Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology, 3*, 371–405.

Language Dynamics in the Framework of Complex Networks: A Case Study on Self-Organization of the Consonant Inventories

Animesh Mukherjee, Monojit Choudhury, Niloy Ganguly, and Anupam Basu

Abstract In this chapter, we present a statistical mechanical model of language acquisition and change at a mesoscopic level, and validate our model for the sound systems of the languages across the world. We show that the emergence of the linguistic diversity that exists across the consonant inventories of some of the major language families of the world can be explained through a complex network based growth model, which has only a single tunable parameter that is meant to introduce a small amount of randomness in the otherwise preferential attachment based growth process. The experiments with this model parameter indicates that the choice of consonants among the languages within a family are far more preferential than it is across the families. Furthermore, our observations indicate that this parameter might bear a correlation with the period of existence of the language families under investigation. These findings lead us to argue that preferential attachment seems to be an appropriate high level abstraction for language acquisition and change.

1 Introduction

Language is a complex physical system. Therefore, like any other complex system its structure and dynamics can be studied at three levels: *microscopic*, *mesoscopic* and *macroscopic*. At a microscopic level language can be viewed as a system emerging from the interactions of socially embedded agents. These agents communicate

A. Mukherjee (✉) · N. Ganguly · A. Basu
Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur 721302, West Bengal, India
e-mail: animeshm@cse.iitkgp.ernet.in; niloy@cse.iitkgp.ernet.in; anupam@cse.iitkgp.ernet.in

M. Choudhury
Microsoft Research India, Bangalore 560025, Karnataka, India
e-mail: monojitc@microsoft.com

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_3, © Springer-Verlag Berlin Heidelberg 2013

through utterances, the collection of which comprise a language at any given time. On the other extreme, at a macroscopic level, a language as a whole can be viewed as a dynamical system defined by its gross properties (e.g., syllable structures and word ordering constraints). Between these two extremes, one can also view a language as a complex system emerging from the interactions between abstract linguistic entities, such as phonemes, morphemes and lexemes. This third view is often referred to as the *mesoscopic* view of a linguistic system. An alternative way to look at this threeway division could be as follows. At the microscopic level, the speakers of a language are the actors. At the macroscopic level, the language as a whole is the actor, while at the mesoscopic level, the linguistic entities are the actors. Note that these three views are not contradictory; rather they describe the properties of the system at different levels.

Any linguistic phenomenon can likewise be studied at all these three levels. In the context of *language acquisition* these three levels could be defined in the following manner. At the microscopic level, we study the process by which human beings (i.e., the speakers) acquire language in a given environment. This, in fact, is the most common interpretation of the term, and, consequently, the most well-studied aspect of language acquisition. Nevertheless, one can also view language acquisition from a macroscopic perspective where the language itself changes over time under certain circumstantial pressures. This phenomenon, which is caused, among many other things, by the language acquisition process at microscopic level, is commonly referred to as *language change*. Although the connection between language change and language acquisition is almost unanimously acknowledged, the nature of this connection is not yet fully understood. This gap can possibly be bridged by studying language acquisition at the mesoscopic level, where the linguistic entities directly interact with each other to give rise to languages. On one hand, the dynamics of their interaction defines the dynamics of language change, and on the other hand, these interactions themselves are the result of underlying language acquisition processes at a microscopic level.

In this chapter, we shall describe some models of language acquisition from a mesoscopic perspective. In particular, we will discuss a series of studies on structure and dynamics of consonant inventories, and how language acquisition could possibly be the underlying force governing this dynamics. Inspired by recent advances in physics of mesoscopy, we employ complex network as a tool to study consonant inventories at a mesoscopic level. While our discussions here will mainly take evidence from the particular case of consonant inventories, our objective here is to show that network based models of linguistic phenomena can aptly model the dynamics of learning at mesoscopy and therefore, provides a promising approach to understand the connections between language acquisition and language change.

In recent times, complex network has been popularly used to model physical, social, biological and cognitive systems (see [35] for a dated but accessible review). A network is a collection of *nodes* representing the entities present in the system and *edges* running between the nodes that represent the interaction patterns between entities. The physical significance of the nodes and the edges depends on the system

being modeled. However, once the network is constructed from real world data, its topological properties provide insights into not only the current organization of the underlying system, but also its evolution, i.e., how this organization might have emerged through simple interactions between the entities over time. This latter part is typically studied through simulations and mathematical analyses of network growth or synthesis models which are stochastic processes capable of explaining the complex structural properties of the network.

There has been some studies on modeling language as a network of linguistic entities (see [13] for a review). In most of these studies, nodes typically represent words, and edges represent various types of word-word interaction (e.g., co-occurrence, or phonological, orthographic, semantic or distributional similarities). Although models of word networks synthesis provide useful insights into cognitive processes beneath the linguistic phenomena, most of these studies do not explicitly relate the network growth model to language acquisition. In this chapter, we shall see that one of the fundamental models of network growth – *preferential attachment* can in fact, be interpreted as a mesoscopic view of language acquisition. Several other learning principles can also be appropriately modeled and studied within the framework of network synthesis.

The rest of the chapter is organized as follows. Section 2 introduces the problem, structure and evolution of phonological inventories, which is used here as the running example to illustrate the new computational framework. In Sect. 3, we present the formal definition of a network-based model of phonological inventories and outline its construction procedure. We analyze some interesting topological properties of this network in the following section. In Sect. 5, we present a synthesis model that can, quite accurately, reproduce the structure of PlaNet. The next section presents an mathematical analysis of the proposed synthesis model and discusses its connection to language acquisition. Section 7 reports further experiments where we analyze the networks constructed for specific language families and discover an interesting connection between the age of a language family and a parameter of our model. This helps us understand the relation between the phenomena of language acquisition and change, and emergence of linguistic diversity and markedness hierarchy. In Sect. 8, we summarize some of the important contributions of this article and outline the scope of future research.

2 Phonological Inventories: A Primer

The most basic units of human languages are the speech sounds. The repertoire of sounds that make up the sound inventory of a language are not chosen arbitrarily, even though the speakers are capable of perceiving and producing a plethora of them. In contrast, the inventories show exceptionally regular patterns across the languages of the world, which is arguably an outcome of the self-organization that goes on in shaping their structure [37]. Earlier researchers have proposed various functional principles such as *maximal perceptual contrast* [29], *ease of*

articulation [15, 29] and *ease of learnability* [15] to explain this self-organizing behavior of the sound inventories. These principles are applied to language as a whole, thereby, viewing it from the macroscopic level. In fact, the organization of the vowel inventories across languages has been quite satisfactorily explained in terms of the single principle of maximal perceptual contrast through linguistic arguments [50], numerical simulations [27,28,43] as well as genetic algorithms [24]. With the advent of highly powerful computers, it has also been possible to model the micro-level dynamics involving a group of (robotic) speakers and their interactions and this in turn has proved to be highly successful in explaining how the vowel inventories originated and self-organized themselves over the linguistic generations [15].

Right from the beginning of the twentieth century, there have been a large number of linguistically motivated attempts [9, 14, 48, 49] in order to explain the emergence of the regularities that are observed across the consonant inventories. However, unlike the case of vowel inventories, the majority of these efforts are limited to the investigation of certain specific properties primarily because of the inherent complexity of the problem. The complexity arises from the fact that (a) consonant inventories are usually much larger in size and are characterized by much more articulatory/acoustic features than the vowel inventories, and (b) no single force is sufficient to explain the organization of these inventories; rather a complex interplay of forces collectively shape their structure. Thus, a versatile modeling methodology, which is hitherto absent in the literature, is required so that the problem can be viewed and solved from an alternative perspective.

Most of the studies on phonological inventories [15,22,25,29] including the ones described here have been conducted on the UCLA Phonological Segment Inventory Database (UPSID) [31]. We have selected UPSID mainly due to two reasons – (a) it is one of the largest database of this type that is currently available and, (b) it has been constructed by selecting languages from moderately distant language families, which ensures a considerable degree of genetic balance.

The languages that are included in UPSID have been chosen in a way to approximate a properly constructed quota rule based on the genetic groupings of the world's extant languages. The quota rule is that only one language may be included from each small language family (e.g., one from the West Germanic and one from the North Germanic) but that each such family should be represented. Eleven major genetic groupings of languages along with several smaller groups have been considered while constructing the database. All these together add up to make a total of 317 languages in UPSID. Note that the availability as well as the quality of the phonological descriptions have been the key factors in determining the language(s) to be included from within a group; however, neither the number of speakers nor the phonological peculiarity of a language has been considered.

Each consonant in UPSID is characterized by a set of articulatory features (i.e., place of articulation, manner of articulation and phonation) that distinguishes it from the other consonants. Certain languages in UPSID also consist of consonants that make use of secondary articulatory features apart from the basic ones. There are around 52 features listed in UPSID; the important ones are noted in Table 1.

Table 1 Some of the important features listed in UPSID

Manner of articulation	Place of articulation	Phonation
tap	velar	voiced
flap	uvular	voiceless
trill	dental	
click	palatal	
nasal	glottal	
plosive	bilabial	
r-sound	alveolar	
fricative	retroflex	
affricate	pharyngeal	
implosive	labial-velar	
approximant	labio-dental	
ejective stop	labial-palatal	
affricated click	dental-palatal	
ejective affricate	dental-alveolar	
ejective fricative	palato-alveolar	
lateral approximant		

Note that in UPSID the features are assumed to be binary-valued (1 meaning the feature is present and 0 meaning it is absent) and therefore, each consonant can be represented by a binary vector.

Over 99 % of the UPSID languages have bilabial (e.g., /p/), dental-alveolar (e.g., /t/) and velar (e.g., /k/) plosives. Furthermore, voiceless plosives outnumber the voiced ones (92 % vs. 67 %). According to [30], languages are most likely to have 8–10 plosives; nevertheless, the scatter is quite wide and only around 29 % of the languages fall within the mentioned limits. Ninety-three percent of the languages have at least one fricative (e.g., /f/). However, as [30] points out, the most likely number of fricatives is between 2 to 4 (around 48 % of the languages fall within this range). Ninety-seven percent of the languages have at least one nasal (e.g., /m/); the most likely range reported in [30] is 2–4 and around 48 % of the languages in UPSID are in this range. In 96 % of the languages there is at least one liquid (e.g., /l/) but, languages most likely have two liquids (around 41 %) [30]. Approximants (e.g., /j/) occur in fewer than 95 % of the languages; however, languages are most likely to have two approximants (around 69 %) [30]. About 61 % of the languages in UPSID have the consonant /h/, which is not included in any of the categories already mentioned above. Some of the most frequent consonants in UPSID are, /p/, /b/, /t/, /d/, /tʃ/, /k/, /g/, /ʔ/, /f/, /s/, /ʃ/, /m/, /n/, /ɲ/, /ŋ/, /w/, /l/, /r/, /j/, /h/, and together they are often termed as the 'modal' inventory [30].

This extremely skewed distribution of consonants across world's languages led many linguists to ask whether there is an inherent bias towards inclusion of certain consonants, such as /m/, in a language. This bias is expressed in phonological theory through a *Markedness hierarchy*, where it is assumed that there is a universal hierarchy of speech sounds (see [4,41] for general discussions on markedness theory and [21] for an introduction to phonological markedness). The sounds lowest in

the markedness hierarchy are preferred by all languages, whereas languages tend to avoid highly marked sounds. While this observation is uninamously accepted based on statistical evidence, there is no single theory for explaining markedness. Some of the popular theories invoke ease of articulation, perception or learning that makes certain sounds more prevelant than others. In other words, markedness hierarchy of the speech sounds have often been explained in terms of phonetic properties of the sounds that make some of these sounds inherently easier to generate, recognize or learn by the human language faculty. Such theories, however, have been criticised based on empirical evidence, which neither supports absolute universality of the phonological markedness hierarchy across languages, nor can it strongly establish the inherent ease of using certain sounds over others. Furthermore, these theories have also attracted philosophical criticisms due to their strong assumptions of inherent biases which may not be necessary at all for explaining skewed distributions.

3 Network Model of Consonant Inventories

A set of consonant inventories can be modeled as a *bipartite network*,[1] where one of the partitions consist of language nodes and the other partition consists of the consonant nodes. Presence of a consonant in a the inventory of a language is represented by an edge connecting the consonant and the language nodes. We shall refer to this network as the **P**honeme-**L**anguage **Net**work or **PlaNet**. In this section we will formally define PlaNet and describe its construction.

3.1 Definition of PlaNet

PlaNet is a bipartite graph $G = \langle V_L, V_C, E_{pl} \rangle$ consisting of two sets of nodes namely, V_L (labeled by the languages) and V_C (labeled by the consonants); E_{pl} is the set of edges running between V_L and V_C. There is an edge $e \in E_{pl}$ from a node $v_l \in V_L$ to a node $v_c \in V_C$ iff the consonant c is present in the inventory of language l. Figure 1 presents a hypothetical example illustrating the nodes and edges of PlaNet.

The representation of the inventories as a bipartite network is motivated by similar modeling of various complex phenomena observed in society as well

[1] A bipartite network is a special kind of network which can be partitioned into two distinct and mutually exclusive sets of nodes such that edges run only between nodes from two different partitions. There are no edges connecting two nodes in the same partition.

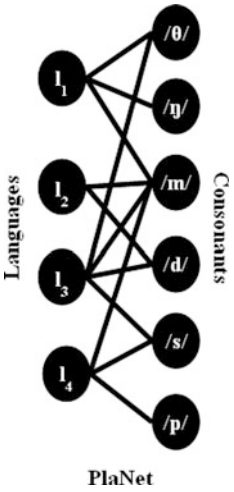

Fig. 1 A hypothetical example illustrating the nodes and edges of PlaNet

as nature, such as (a) the movie-actor network [2, 3, 38, 42, 51] where movies and actors constitute the two respective partitions and an edge between them signifies that a particular actor acted in a particular movie, (b) the article-author network [7, 26, 34] where the two partitions respectively correspond to articles and authors and edges denote which person has authored which articles and (c) the board-director network [12, 47] where the two partitions correspond to the boards and the directors respectively and a director is linked by an edge with a society if he/she sits on its board. In fact, the concept of bipartite networks has also been extended to model such diverse phenomena as the city-people network [17] where an edge between a person and a city indicates that he/she has visited that city, the word-sentence network [19, 20], the bank-company network [46] or the donor-acceptor network [45].

3.2 Construction Methodology

We have used UPSID in order to construct PlaNet. Consequently, the total number of language nodes in PlaNet (i.e., $| V_L |$) is 317. The total number of distinct consonants found across the 317 languages of UPSID, after appropriately filtering the *anomalous* and the *ambiguous* ones [31], is 541. In UPSID, a phoneme has been classified as anomalous if its existence is doubtful and ambiguous if there is insufficient information about the phoneme. For example, the presence of both the palatalized dental plosive and the palatalized alveolar plosive are represented in UPSID as palatalized dental-alveolar plosive (an ambiguous phoneme). According to popular techniques [39], we have completely ignored the anomalous phonemes from the data set, and included all the ambiguous forms of a phoneme as separate

phonemes because, there are no descriptive sources explaining how such ambiguities might be resolved. Therefore, the total number of consonant nodes in PlaNet (i.e., $| V_C |$) is 541.

The number of edges in PlaNet (i.e., $| E_{pl} |$) is 7,022. Thus, the connection density of PlaNet is $\frac{|E_{pl}|}{|V_L||V_C|} = \frac{7{,}022}{317 \times 541} = 0.06$, which can also be thought of as the probability that a randomly chosen consonant occurs in a particular language. However, as we shall see below, the occurrence of the consonants does not depend upon a single probability value; rather, it is governed by a well-behaved probability distribution.

4 Topological Properties of PlaNet

In this section, we shall study the topological properties of PlaNet mainly in terms of the degree distributions of its two sets of nodes.

4.1 Degree Distribution of PlaNet

The degree of a node v, denoted by k_v, is the number of edges incident on v. Therefore, the degree of a language node v_l in PlaNet refers to the size of the consonant inventory of the language l. Similarly, the degree of a consonant node v_c in PlaNet refers to the frequency of occurrence of the consonant c across the languages of UPSID.

The degree distribution is the fraction of nodes, denoted by p_k, that have a degree equal to k [35]. In other words, it is the probability that a node chosen uniformly at random from the network (with N nodes) has a degree equal to k. The cumulative degree distribution P_k is the fraction of nodes having degree greater than or equal to k. Therefore,

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{1}$$

Note that the cumulative distribution is more robust to noise present in the observed data points, but at the same time it contains all the information encoded by p_k [35].

4.1.1 Degree Distribution of the Language Nodes

Figure 2 shows the degree distribution of the nodes in V_L where the x-axis denotes the degree of each language node expressed as a fraction of the maximum degree and the y-axis denotes the fraction of nodes having a given degree.

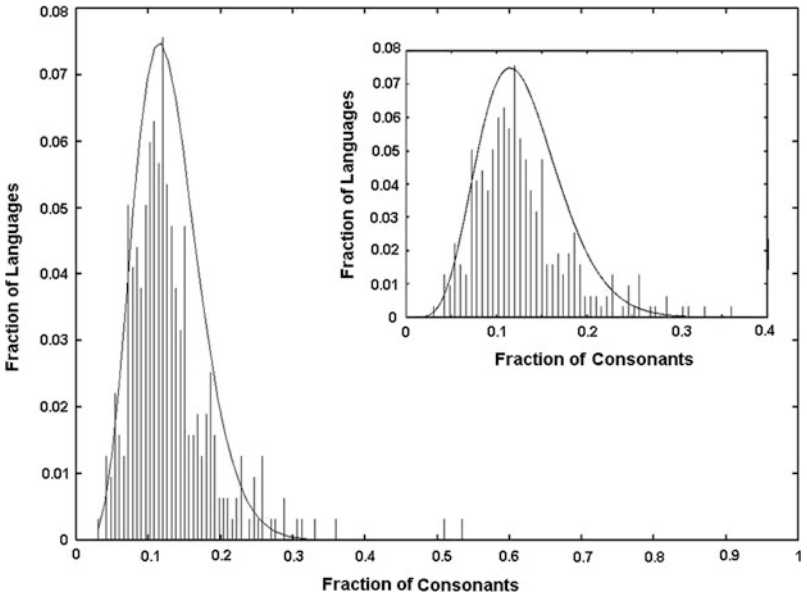

Fig. 2 Degree distribution of the language nodes in PlaNet. The figure in the inset is a magnified version of a portion of the original figure

Figure 2 indicates that the number of consonants appearing in different languages follow a β-distribution[2] (see [10] for reference) which is right skewed with the values of α and β equal to 7.06 and 47.64 (obtained using maximum likelihood estimation method) respectively. This asymmetry in the distribution points to the fact that languages usually tend to have smaller consonant inventory size, the best value being somewhere between 10 and 30. The distribution peaks roughly at 21 (which is its mode) while the mean of the distribution is also approximately 21 indicating that on an average the languages in UPSID have a consonant inventory of size 21 (approx.) [32].

4.1.2 Degree Distribution of the Consonant Nodes

Figure 3 illustrates the cumulative degree distribution plot for the consonant nodes in V_C in doubly-logarithmic scale. In this figure the x-axis represents the degree k and the y-axis represents the distribution P_k.

[2] A random variable is said to have a β-distribution with parameters $\alpha > 0$ and $\beta > 0$ if and only if its probability mass function is given by, $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$ and $f(x) = 0$ otherwise. $\Gamma(\cdot)$ is the Euler's gamma function.

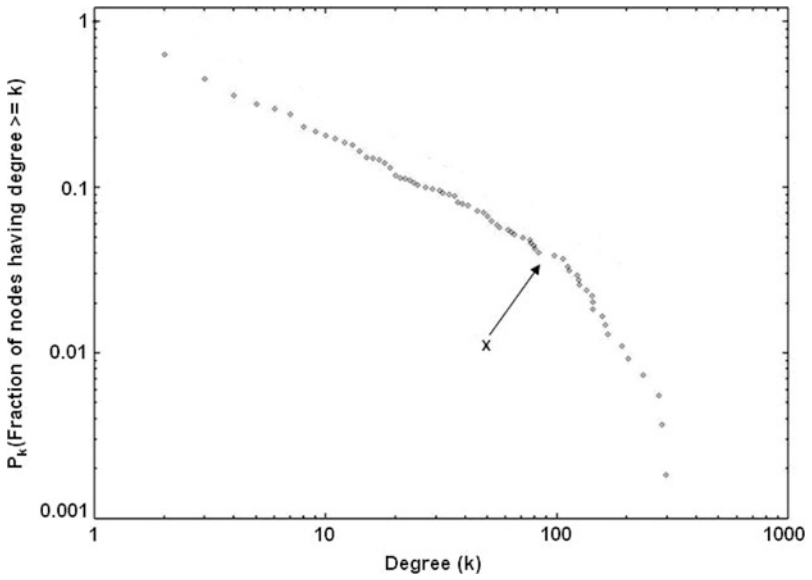

Fig. 3 Degree distribution of the consonant nodes in PlaNet in doubly-logarithmic scale. The letter **x** denotes the cut-off point

As can be seen from Fig. 3, P_k follows a power-law with an exponential cut-off. Sometimes such distributions have also been referred to as two-regime power-law. The cut-off point is shown by the letter **x** in the figure. We find that there are 22 consonant nodes which have their degree above the cut-off range (i.e., these are the extremely frequent consonants). It is worth mentioning that power-law distribution of word frequencies in text corpora is a very well studied phenomenon and is commonly referred to as Zipf's law. Power-law distributions are very skewed; in the specific context of PLaNet, we can interpret it as a hierarchical arrangement of the consonants where a few consonants that are at the top of these hierarchy are present in almost all the languages of the world; whereas, a large number of consonants are at the lower ends of these hierarchy and are present only in a handful of world's languages.

Recall that the skewed disrtibution of consonants over languages is a well-known fact, which linguists refer to as the markedness hierarchy. However, a network based representation immediately allows us to seek for an evolutionary model that can explain the network topology and therefore, the degree distribution. Earlier studies have shown that in most of the networked systems like the society, the Internet and the World Wide Web, *preferential attachment* (i.e., when "the rich gets richer") [6, 44] is known to play a crucial role in generating power-law degree distributions. Therefore, in the following section, we will explore a preferential attachment based synthesis model to explain the evolution of PLaNet.

Algorithm 1: Synthesis model based on preferential attachment

Input: Nodes L_1 through L_{317} sorted in an increasing order of their degrees

for t = *1 to 317* **do**
  Choose (in order) a node L_j with degree d_j;
  for c = *1 to* d_j **do**
    Connect L_j to a node $C_i \in V_C$ to which it is not already connected following the distribution, $Pr(C_i) = \frac{\gamma k_i + 1}{\sum_{i' \in V_C'} (\gamma k_{i'} + 1)}$ where V_C' is the set of nodes in V_C (inclusive of C_i) to which L_j is not yet connected, k_i is the current degree of node C_i and γ is the tunable parameter;
  end
end

5 The Synthesis Model

Let us assume that the distribution of the consonant inventory size, i.e., the degrees of the language nodes is known a priori. Let the degree of a language node $L_i \in V_L$ be denoted by d_i. The consonant nodes in V_C are assumed to be unlabeled, i.e., they are not marked by the distinctive features that characterize them. We first sort the nodes L_1 through L_{317} in the ascending order of their degrees. At each time step a node L_j, chosen in order, preferentially attaches itself with d_j *distinct* nodes (call each such node C_i) of the set V_C. The probability $Pr(C_i)$ with which the node L_j attaches itself to the node C_i is given by,

$$Pr(C_i) = \frac{\gamma k_i + 1}{\sum_{i' \in V_C'} (\gamma k_{i'} + 1)} \tag{2}$$

where, k_i is the current degree of the node C_i, V_C' is the set of nodes in V_C that are not already connected to L_j and γ is a positive tunable parameter that controls the amount of randomness in the system. The lower the value of γ the higher is the randomness. For instance, at $\gamma = 0$, the attachment probability $Pr(C_i)$ is equal to $1/N$ (N being the total number of consonants in the system), regardless of the degree of the node. Thus, instead of a degree based preferential attachment, at $\gamma = 0$ the model boils down to a random attachment model, which is known to generate multinomial degree distributions. On the other hand, if γ is very large, the probaility of attachment becomes purely preferntial. Note that the positive constant $1/\gamma$ is usually referred to as the *initial attractiveness* [16], because higher the value of this parameter, higher is the probability that a node with 0 degree (no incoming edges) will get a new connection.

Algorithm 1 summarizes the mechanism to generate the synthesized version of PlaNet (henceforth PlaNet$_{syn}$) and Fig. 4 illustrates a partial step of the synthesis process. In the figure, when language l_4 has to connect itself with one of the nodes in the set V_C it does so with the one having the highest degree (=3) rather than with

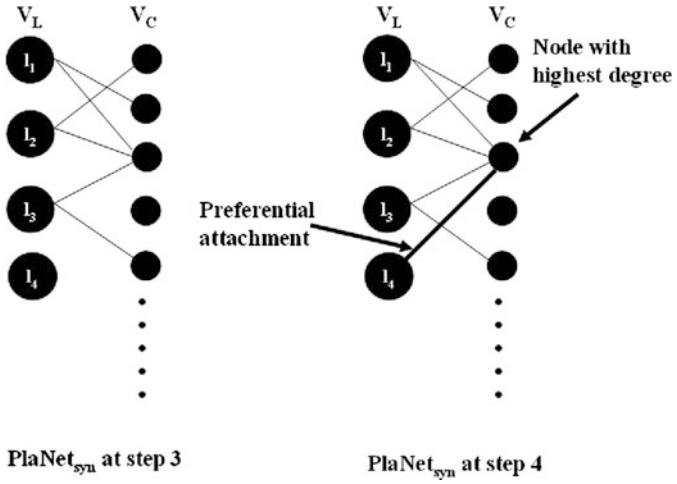

Fig. 4 A partial step of the synthesis process

others in order to achieve preferential attachment which is the working principle of ouralgorithm.

We simulate the above model to obtain PlaNet_{syn} for 100 different runs and average the results over all of them. We find that the degree distributions that emerge fit the empirical data well for $\gamma \in [12.5,16.7]$, the best being at $\gamma = 14$ (shown in Fig. 5). In fact, the mean error[3] between the real and the synthesized distributions for $\gamma = 14$ is as small as 0.03. In contrast, if there is no preferential attachment and all the connections to the consonant nodes are equiprobable (see Fig. 5), then this error rises to 0.35.

Apart from the ascending order, we have also simulated the model with descending and random order of the inventory size. The degree distribution obtained by considering the ascending order of the inventory size matches much more accurately than in the other two scenarios. One possible reason for this might be as follows. Each consonant is associated with two different frequencies: (a) the frequency of occurrence of a consonant over languages or the *type* frequency, and (b) the frequency of usage of the consonant in a particular language or the *token* frequency. Researchers have shown in the past that these two frequencies are positively correlated [11]. Nevertheless, our synthesis model based on preferential attachment takes into account only the type frequency of a consonant and not its token frequency. If language is considered to be an evolving system then both of

[3]Mean error is defined as the average difference between the ordinate pairs (say y and y') where the abscissas are equal. In other words, if there are Y such ordinate pairs then the mean error can be expressed as $\frac{\sum|y-y'|}{Y}$.

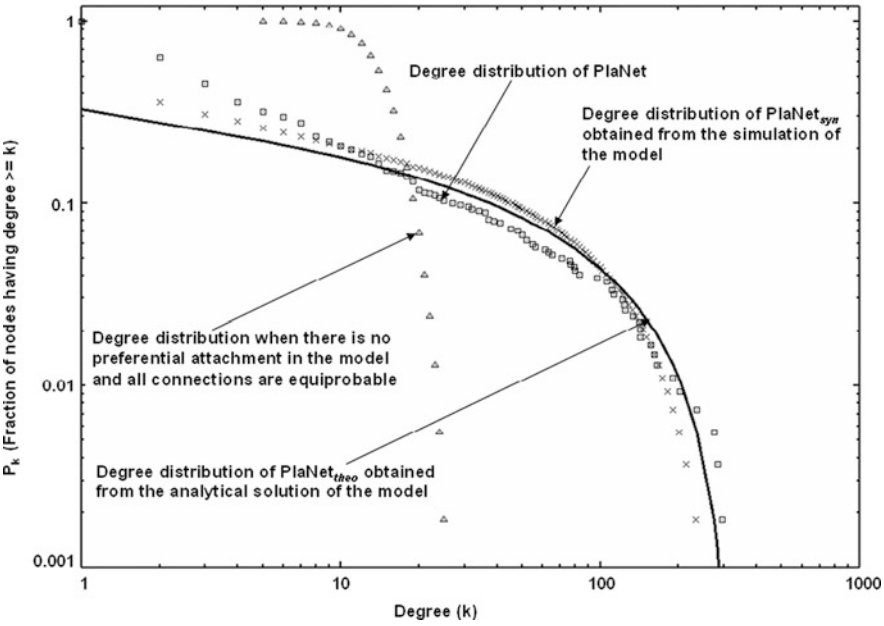

Fig. 5 Comparison of the degree distribution (in doubly-logarithmic scale) of the consonant nodes in PlaNet with that of (*a*) PlaNet$_{syn}$ obtained from the simulation of the model ($\gamma = 14$) and (*b*) PlaNet$_{theo}$ obtained from the analytical solution of the model ($\gamma = 14$). The results are also compared with the case where there is no preferential attachment and all the connections are equiprobable

these frequencies, in one generation, should play an important role in shaping the inventory structure of the next generation.

In the later stages of our synthesis process when the attachments are strongly preferential, the type frequencies span over a large range and automatically compensate for the absence of the token frequencies (since they are positively correlated). However, in the initial stages of this process the attachments that take place are random in nature and therefore, the type frequencies of all the nodes are roughly equal. At this point it is the token frequency (absent in our model) that should discriminate between the nodes. This error due to the loss of information of the token frequency in the initial steps of the synthesis process can be minimized by allowing only a small number of attachments (so that there is less spreading of the error). This is perhaps the reason why sorting the language nodes in the ascending order of their degree helps in obtaining better results.

Note that even though we make a strong assumption that the degrees of the language nodes are known a priori, it is neither necessary nor a strong limitation of our experiments and their conclusion. Alternative models where the degree of all the language nodes is same and equal to the mean of the distribution, or are sampled from the degree distribution of language nodes (Fig. 2) also lead to very

similar network topologies. In other words, regardless of the degrees of the language nodes the proposed synthesis model is able to explain the two-regime power-law distribution of the consonant degrees.[4] However, for understandable reasons, the degree distribution of the synthesized networks are quantitatively closest to that of real PlaNet when this assumption is made.

6 Interpretation of the Synthesis Model

In this section we ask two important questions about the proposed synthesis model. First, what is the mathematical interpretation of the model? In other words, we would like to analyze the model and theoretically predict the nature of emergent degree distribution under various parameter settings. Second, what is the physical interpretation of the model? Stated differently, why or how this model is able to explain the distribution of the consonants over languages? How is it, if at all, related to language acquisition and change.

6.1 Mathematical Analysis of the Model

Several models have been proposed in literature to synthesize the structure of these bipartite networks, i.e., when both the partitions grow unboundedly with time [2,20,38,42,51]. The results of such growth models indicate that when an incoming movie node (in case of movie-actor networks) *preferentially* attaches itself to an actor node, the emergent degree distribution of the actor nodes follows a power-law (see [42] for details). This result is reminiscent of unipartite networks where *preferential attachment* leads to the emergence of power-law degree distributions (see [6] for details).

In most of these networks, however, both the partitions grow with time unlike PlaNet where the partition corresponding to the consonants remains relatively fixed over time while the partition corresponding to the languages grows with time. Although there have been some studies on non-growing bipartite networks [18,36], those like PlaNet where one of the partitions remain fixed over time (i.e., the partition of consonants) while the other grows (i.e., the partition of languages) have received much less attention. It is difficult to solve this model because, unlike the popular preferential attachment based synthesis models for unipartite [6] and bipartite [42] networks, in this case, one cannot make the stationary state assumption $p_{k,t+1} = p_{k,t}$ in the limit $t \to \infty$ (here, $p_{k,t}$ denote the probability that a randomly

[4] As one can see in Fig. 5, the analytical solution arrived at by assuming that all language nodes have the same degree μ is qualitatively similar to the degree distribution of real PlaNet, though a poorer match to it when we comapre the simulation with exact degrees of the language nodes.

chosen node from the partition V_C has degree k after t time steps). This is due to the fact that the average degree of the nodes in V_C diverges with time and consequently, the system does not have a stationary state.

Nevertheless, for certain simplifications of the model we can derive an approximate closed form expression for the degree distribution of the V_C partition after a given time step t. More specifically, we assume that the degree of the nodes in the V_L partition is equivalent to their average degree and is therefore, a constant (μ). In other words, μ represents the average size of a consonant inventory or the average number of consonants present in human languages. We further assume that in a time step a language node can attach itself to a consonant node more than once. Although by definition, a consonant can occur in the inventory of a language only once, as we shall see, the result derived with the above assumption matches fairly well with the empirical data.

Under the assumptions mentioned above the denominator of the Eq. 2 can be re-written as $\sum_{i=1}^{N}(\gamma k_i + 1)$ where $N = \mid V_C \mid$. Further, since the sum of the degrees in the V_L partition after t steps ($= \mu t$) should be equivalent to that in the V_C partition therefore we have

$$\sum_{i=1}^{N} k_i = \mu t \tag{3}$$

Notice that the average degree of the nodes in V_C after t steps is $\mu t/N$ which, as we have pointed out earlier, diverges with t because, N is fixed over time.

At time $t = 0$, all the nodes in V_C have a degree zero and therefore our initial condition is $p_{k,t=0} = \delta_{k,0}$, where $\delta_{k,0}$ is the Kronecker delta function [23]. After time t, the degree distribution of the consonant nodes is given by the following equation:

$$p_{k,t} \approx= \widehat{A}(t, \eta, \gamma)(k/t)^{\gamma^{-1}-1}(1 - k/t)^{\eta-\gamma^{-1}-1} \tag{4}$$

where $\eta = N/\mu\gamma$, and

$$\widetilde{A}(t, \gamma, \eta) = \frac{t^{t+0.5}\eta^{\eta+0.5}\gamma^{\gamma^{-1}-0.5}e}{\sqrt{2\pi}(t+\eta)^{t+\eta+0.5}(\eta-\gamma^{-1})^{\eta-\gamma^{-1}+0.5}} \tag{5}$$

Refer to the Appendix for a detailed derivation of the above formulae. A concise version of this solution has been presented in [40]. We shall refer to this analytically derived solution of PlaNet as PlaNet$_{theo}$.

Recall that in Sect. 5 we have found through simulations that the best fit for the degree distribution emerges at $\gamma = 14$. Replacing μ by 21, t by 317, N by 541 and γ by 14 we obtain the degree distribution for the consonant nodes $P_{k,t}$ of PlaNet$_{theo}$. The bold line in Fig. 5 illustrates the plot for this distribution in doubly-logarithmic scale. The figure indicates that the theoretical curve (i.e., the degree distribution of PlaNet$_{theo}$) matches quite well with the empirical data (i.e., the degree distribution of PlaNet). In fact, the mean error between the two curves in this case is as small as 0.03. It is worthwhile to mention here that since the degree distribution obtained from the simulation as well as the theoretical analysis of the model matches the

real data for a very high value of γ there is a considerable amount of preferential attachment that goes on in shaping the emergent structure of PlaNet.

There are two important observations that one can make from the analysis and the results. First, unlike the case of bipartite networks where both the partitions grow, in our model of PLaNet we observe a Beta distribution rather than a power-law. By controlling the parameter γ, one can change the skewness of the distribution and make it look more like a power-law on one extreme and a normal-like distribution on another extreme. See [40] for a detailed description of the qualitatively different distributions that can emerge in this model. Second, it turns out that for consonant inventories the value of γ is fairly high that gives rise to a very skewed distribution, but not high enough to result in a "winner takes all" situation.[5]

6.2 Linguistic Interpretation of the Model

There are several ways in which one can interpret preferential attachment and the role of γ in the context of evolution of consonant inventories. Arrival of a new language in the system can be interpreted as a new language being created through the process of language change. This could be either due to contact between two or more languages, or due to some kind of drift from an existing mother language. In either case, the predecessor(s) of the new language already exist(s) in the system. Therefore, it's consonants are also present in the system with non-zero degree. There is a very high chance that the new language will inherit the consonants of its predecessor(s) with some minor variations, if any. Therefore, the chances of the new language node connecting to those consonants which are already connected to other nodes in the system is higher than its probability to connect to a consonant node with degree 0. Under the assumption that to start with there were very few mother languages in the system, one can show that consonants that were more present in those mother languages will eventually become very widespread across all the languages, while those which were not present initially will have very low probability of incoming edges in the future. In other words, consonants that are more prevalent now in the system will eventually become even more prevelant in the future. This, precisely, is the crux of the preferential attachment model.

According to this interpretation $1/\gamma$ reflects the probability that during the process of language change a language acquires new consonants distinct from those present in its predecessor(s). In reality and also in our model this probability is greater than 0, though quite small. One interesting prediction borne out of this interpretation is as follows. Since PLaNet has been constructed from the phonological inventories of diverse language families and since we do not specifically seed the

[5] As explained in [40], when $\gamma \geq (N/\mu) - 1$ all the almost all the language nodes connect to the same set μ consonant nodes making other consonants virtually inexistant. For such a situation to arise for PLaNet, γ had to be greater than or equal to 25.

model with the set of mother languages of all the families, the value of γ estimated from PLaNet will be lower (i.e., it will reflect a higher probability of choosing a new consonant during the process of language change) than its real counterpart. One way to test this hypothesis is to construct PLaNets with languages from a single family and estimate the value of γ for them. Our interpretation suggests that the value of γ's for family-specific PLaNets will be higher than 14. Furthermore, greater contact with languages from other families should ideally lead to more diversity in the inventories of a language family leading to a lower value of γ. In the next section, we will describe some experiments that empirically justify these intuitions.

Yet another interpretation of the proposed model could be based on language acquisition. Consonants belonging to languages that are more prevalent among the speakers in one generation have higher chances of being transmitted to the speakers of languages of the subsequent generations than other consonants (see [8] for similar observations). Recall that there is a very high correlation between type and token frequencies of consonants. Therefore, consonants which are more prevelant across language at a given time are also usually more commonly used in the languages which have these consonants. Imagine a new speaker, i.e., a child entering the system. The probability of the child acquiring a consonant is directly proportionate to the amount of its exposure received by the child. This exposure, in turn, is proportionate to the current usage of the consonant in the linguistic community the child resides in. Thus, higher the usage of a consonant at a particular point of time, higher is the probability that learners will acquire and use this consonant. In other words, more prevalent a consonant is in the environment today, the more prevalent it will be in the future. This line of argument again explains the manifestation of preferential attachment in linguistic system. In this learning based interpretation of our model, γ represents fidelity of language acquisition. A higher value of γ would mean that a child acquires the language of his/her parents and the environment without any distortion, whereas lower values of γ would indicate that language acquisition is a noisy process and children can learn languages very different from what they have ever been exposed to. An extreme case is when $\gamma = 0$, where the child acquires a random language irrespective of the stimulus present in the environment.

The two interpretations of the model described above are at two different levels. While the language change based interpretation connects the mesoscopic evolutionary dynamics of the network to macroscopic processes, language acquisition based interpretation links it to the microscopic dynamics of language acquisition. Note that our interpretations are not specific to phonological systems. In fact, they can be applied to any system which evolves through ontogenic transmission of knowledge. Since several, if not all, aspects of a linguistic system are transmitted through language acquisition, we believe that preferential attachment based models at a mesoscopic level will be able to explain distributional patterns of linguistic entities. Indeed, several other linguistic units such as words, syllables, syntactic structures and lexemes are also known to follow Zipfian or very skewed distributions![52]. Recall that linguists often invoke theory of markedness to explain such universally prevalent skewed distributions of linguistic units. However, our experiments and

analysis show that it is not necessary to invoke an extraneous principle to explain such distributions. Any initial heterogenity in the distribution will always eventually get magnified by several folds due to the nature of the processes involved in language acquisition and change. The initial skew in the distribution can, in fact, be negligibly small and just an effect of random perturbations rather than some inherent biases of the language faculty.

Nevertheless, one should keep in mind that our analysis does not provide any evidence against the theory of markedness. It just says that for explaining skewed statistical distributions of linguistic entities it is not necessary to assume a markedness hierarchy. Apart from explaining skewed distributions, theory of markedness has other roles in linguistic analyses and there are independent ways of verifying those theories.

7 Dynamics of the Language Families

In this section, we investigate the dynamics within and across the consonant inventories of some of the major language families of the world. More specifically, for our investigation, we choose five different families namely the Indo-European, the Afro-Asiatic, the Niger-Congo, the Austronesian and the Sino-Tibetan. We manually sort the languages of these five groups from the data available in UPSID. Note that we have included a language in any group if and only if we could find a direct evidence of its presence in the corresponding family. We next present a brief description of each of these groups[6] and list the languages from UPSID that are found within them.

Indo-European: This family includes most of the major languages of Europe and south, central and south-west Asia. Currently, it has around three billion native speakers, which is largest among all the recognized families of languages in the world. The total number of languages appearing in this family is 449. The earliest evidences of the Indo-European languages have been found to date 4,000 years back.

Languages: Albanian, Bengali, Breton, Bulgarian, Farsi, French, German, Greek, Hindi/Urdu, Irish, Kashmiri, Kurdish, Lithuanian, Norwegian, Pashto, Romanian, Russian, Sinhalese, Spanish.[7]

Afro-Asiatic: Afro-Asiatic languages have about 200 million native speakers spread over north, east, west, central and south-west Africa. This family is divided into five subgroups with a total of 375 languages. The proto-language of this family began to diverge into separate branches approximately 6,000 years ago.

[6]Most of the information has been collected from the Ethnologue: http://www.ethnologue.com/ and the World Atlas of Language Structures: http://wals.info/

[7]Interestingly, while preparing this set of Indo-European languages from UPSID, we did not find English.

Table 2 Number of nodes and edges in the four bipartite networks corresponding to the four language families

Networks	$\|V_L\|$	$\|V_C\|$	$\|E_{pl}\|$	γ	Age (in years)
IE-PlaNet	19	148	534	18.0	4,000 (or 8,000)
AA-PlaNet	17	123	453	26.0	6,000
ST-PlaNet	9	71	201	28.6	6,000
NC-PlaNet	30	135	692	28.6	5,000
AN-PlaNet	12	82	221	33.3	4,000

Languages: Amharic, Angas, Arabic, Awiya, Dera, Dizi, Hamer, Hausa, Iraqw, Kanakuru, Kefa, Kullo, Margi, Ngizim, Shilha, Socotri, Somali.

Niger-Congo: The majority of the languages that belong to this family are found in the sub-Saharan parts of Africa. The number of native speakers is around 300 million and the total number of languages is 1,514. This family descends from a proto-language, which dates back 5,000 years.

Languages: Akan, Amo, Bambara, Bariba, Beembe, Birom, Bisa, Cham, Dagbani, Dan, Diola, Doayo, Efik, Ga, Gbeya, Igbo, Ik, Kadugli, Koma, Kpelle, Lelemi, Moro, Senadi, Tampulma, Tarok, Teke, Temne, Wolof, Zande, Zulu.

Austronesian: The languages of the Austronesian family are widely dispersed throughout the islands of south-east Asia and the Pacific. There are 1,268 languages in this family, which are spoken by a population of six million native speakers. Around 4,000 years back it separated out from its ancestral branch.

Languages: Adzera, Batak, Chamorro, Hawaiian, Iai, Javanese, Kaliai, Malagasy, Roro, Rukai, Tsou, Tagalog.

Sino-Tibetan: Most of the languages in this family are distributed over the entire east Asia. With a population of around two billion native speakers it ranks second after Indo-European. The total number of languages in this family is 403. Some of the first evidences of this family can be traced 6,000 years back.

Languages: Ao, Burmese, Dafla, Hakka, Jingpho, Karen, Lahu, Mandarin, Taishan.

We use the consonant inventories of the language families listed above to construct five bipartite networks – IE-PlaNet (for Indo-European family), AA-PlaNet (for Afro-Asiatic family), NC-PlaNet (for Niger-Congo family), AN-PlaNet (for Austronesian family) and ST-PlaNet (for Sino-Tibetan family). The number of nodes and edges in each of these networks are noted in Table 2.

We attempt to fit the degree distribution of the five empirical networks with the analytical expression derived for $P_{k,t}$ in the previous section. For all the experiments, we set $N = 541$, $t =$ number of languages in the family under investigation and $\mu =$ average degree of the language nodes in the PlaNet representing the family under investigation. Therefore, given the value of k we can compute $p_{k,t}$ and consequently, $P_{k,t}$, if γ is known. We vary the value of γ such that the mean error between the degree distribution of the real network and the equation is minimum. The best fits obtained for each of the five networks are shown in Fig. 6. The values of γ corresponding to these fits are noted in Table 2.

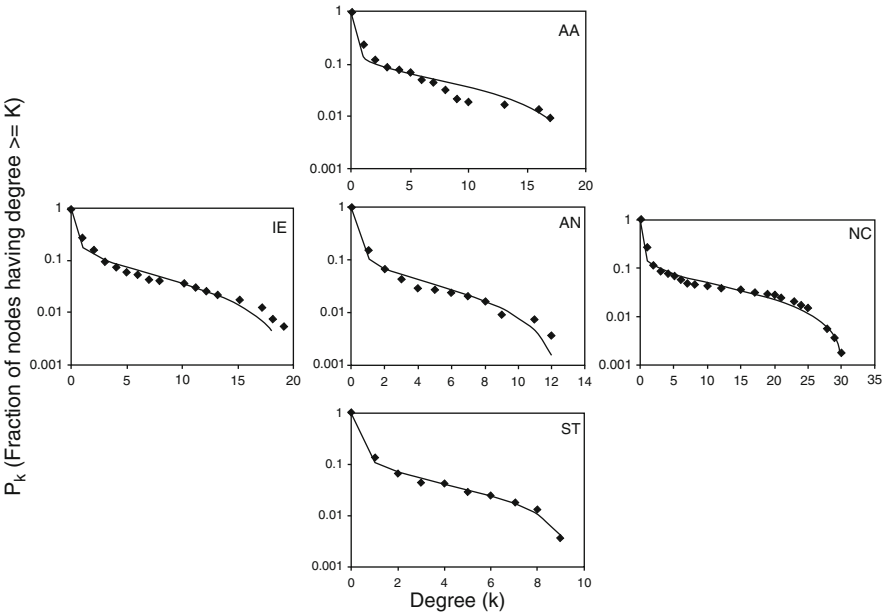

Fig. 6 The degree distribution of the different real networks (*black dots*) along with the best fits obtained from the analytical expression for $P_{k,t}$ (*grey lines*). For all the plots the y-axis is in log-scale

The results indicate that the value of γ for PlaNet is lower than that of all the individual networks corresponding to the language families. Therefore, it may be argued that the preferential component within a language family is stronger than across families. Note that this is true only for real linguistic families and not for any arbitrary group of languages. In fact, if one randomly selects a set of inventories to represent a family then for a large number of such sets the average value of γ is 14.7 which is close to that of PlaNet. These observations neatly concur with our predictions discussed in the last section.

We further observe a very interesting positive correlation between the approximate age of the language family and the values of γ obtained in each case (see Table 2). The only anomaly is the Indo-European branch, which possibly indicates that this might be much older than it is believed to be. In fact, a recent study [5] has argued that the age of this family dates back to 8,000 years. If this last argument is assumed to be true then the values of γ have a one-to-one correspondence with the approximate period of existence of the language families. As a matter of fact, this correlation can be intuitively justified – higher is the period of existence of a family higher are the chances of its diversification into smaller subgroups and contact with languages from other families, which in turn increases the randomness of the system and therefore, the values of γ are found to be less for the older families.

8 Conclusion

In this chapter, we have described a complex network based framework to study the dynamics of linguistic systems and associated phenomena. Unlike the popular studies of language dynamics, which either takes a microscopic view of language acquisition by an individual or a macroscopic perspective on how language as a whole changes over time, here we discuss a mesoscopic framework, where

- Languages and/or linguistic units are modeled as an interacting set of entities represented through nodes and edges of a complex network.
- The topological properties of the network constructed from empirical data are studied in order to understand the principles governing the linguistic phenomena being studied.
- A network synthesis model is then invented, which can potentially replicate all the topological characteristics of the real linguistic network(s). The synthesis model is usually validated through simulation experiments, and if possible also through analytical reasoning.
- Finally, through mathematical analysis and linguistic arguments the general properties and expressiveness of the synthesis model and its parameters are is interpreted in the context of language dynamics. This might also lead to new linguistic predictions that can be verified against real data, whenever available.

Usually, a mesoscopic model so defined in terms of a complex network and a stochastic synthesis process provides useful insights about the link between language acquisition and language change. Furthermore, it is also capable of quantifying the rate of these processes and their affect on a linguistic system (e.g., γ in our case study). Thus, this line of research seems to hold a lot of promise in furthering our understanding of language dynamics.

The current case study on self-organization of phonological inventories spells out only the few initial steps of research in this direction. In fact, the studies described in this chapter could have been carried out even without conceptualizing the underlying network model, though it goes without saying that the network model provides us with a strong visualization of the system and the processes. Nevertheless, topological characteristics of a network extend much beyond degree distribution. While the synthesis model described here is capable of explaining the degree distribution of the network, it cannot explain several other topological characteristics such as the structure of one-mode projection of the network or clustering coefficient. Through a series of systematic experiments, we have shown [33] that a synthesis model based on a linear combination of preferential attachment and economy of distinctive features can explain most of the topological properties of PLaNet. Our work shows that while preferntial attachment seems to be a strong driving force at a global scale, at the level of an individual (or a single language) acquisition of a new consonant is much easier when all of its distinctive features are already known to the individual (or language). Thus, feature economy can sometime make acquisition of rare consonants easier than that of a frequent consonant using a set of features alien to the learner (or language).

Although research in linguistic network has gained much popularity in the recent times, it is still confined within few research communities. Mesoscopic models of language dynamics, and more specifically language acquisition, are rare. On the other hand, there are tremendous research opportunities in this area. For instance, we do not know how words are acquired in the context of other words; how the network of words in the mind changes and evolves over time; how, if at all, language universals are borne out of language dynamics.

Appendix: Derivation of the Analytical Solution

We shall solve the model for $\mu = 1$ and then generalize for the case $\mu > 1$. Please refer to Sects. 5 and 6 for explanation of the model and the notations used.

Solution for $\mu = 1$

Since $\mu = 1$, at each time step a node in the V_L partition essentially brings a single incoming edge as it enters the system. The evolution of $p_{k,t}$ can be expressed as

$$p_{k,t+1} = (1 - \widetilde{P}(k,t))\, p_{k,t} + \widetilde{P}(k-1,t)\, p_{k-1,t} \tag{6}$$

where $\widetilde{P}(k,t)$ refers to the probability that the incoming edge lands on a consonant node of degree k at time t. $\widetilde{P}(k,t)$ can be easily derived for $\mu = 1$ using together the Eqs. 2 and 3 and takes the form

$$\widetilde{P}(k,t) = \begin{cases} \frac{\gamma k+1}{\gamma t+N} & \text{for} \quad 0 \le k \le t \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

for $t > 0$ while for $t = 0$, $\widetilde{P}(k,t) = \frac{1}{N}\delta_{k,0}$.

Equation 6 can be explained as follows. The probability of finding a consonant node with degree k at time $t+1$ decreases due to those nodes, which have a degree k at time t and receive an edge at time $t+1$ therefore acquiring degree $k+1$, i.e., $\widetilde{P}(k,t)\, p_{k,t}$. Similarly, this probability increases due to those nodes that at time t have degree $k-1$ and receive an edge at time $t+1$ to have a degree k, i.e., $\widetilde{P}(k-1,t)\, p_{k-1,t}$. Hence, the net increase in the value of $p_{k,t+1}$ can be expressed by the Eq. 6.

In order to have an exact analytical solution of the Eq. 6 we express it as a product of matrices

$$\mathbf{p}_{t+1} = \mathbf{M}_t \mathbf{p}_t = \Big[\prod_{\tau=0}^{t} \mathbf{M}_\tau\Big]\mathbf{p}_0 \tag{8}$$

where $\mathbf{p}_t$ denotes the degree distribution at time t and is defined as $\mathbf{p}_t = [p_{0,t}\ p_{1,t}\ p_{2,t}\ \ldots]^T$ (T stands for the standard transpose notation for a matrix), $\mathbf{p}_0$ is the initial condition expressed as $\mathbf{p}_0 = [1\ 0\ 0\ \ldots]^T$ and $\mathbf{M}_\tau$ is the evolution matrix at time τ which is defined as

$$\mathbf{M}_{\emptyset} = \begin{pmatrix} 1-\widetilde{P}(0,\tau) & 0 & 0 & 0\ldots \\ \widetilde{P}(0,\tau) & 1-\widetilde{P}(1,\tau) & 0 & 0\ldots \\ 0 & \widetilde{P}(1,\tau) & 1-\widetilde{P}(2,\tau) & 0\ldots \\ \vdots & \vdots & \vdots & \vdots\ddots \end{pmatrix} \tag{9}$$

Let us further define a matrix $\mathbf{H}_t$ as follows.

$$\mathbf{H}_0 = \mathbf{M}_0 \tag{10}$$

$$\mathbf{H}_t = \mathbf{M}_t\mathbf{H}_{t-1} = \Big[\prod_{\tau=0}^{t}\mathbf{M}_\tau\Big] \tag{11}$$

Thus we have,

$$\mathbf{p}_{t+1} = \mathbf{H}_t\mathbf{p}_0 \tag{12}$$

Since our initial condition (i.e., $\mathbf{p}_0$) is a matrix of zeros at all positions except the first row therefore, all the relevant information about the degree distribution of the consonant nodes is encoded by the first column of the matrix $\mathbf{H}_t$. The $(k+1)$th element of this column essentially corresponds to $p_{k,t}$. Let the entry corresponding to the ith row and the jth column of $\mathbf{H}_t$ and $\mathbf{M}_t$ be denoted by $h^t_{i,j}$ and $m^t_{i,j}$ respectively. On successive expansion of $\mathbf{H}_t$ using the recursive definition provided in Eq. 11, we get (see Fig. 7 for an example)

$$h^t_{i,j} = m^t_{i,i-1}h^{t-1}_{i-1,j} + m^t_{i,i}h^{t-1}_{i,j} \tag{13}$$

or,

$$h^t_{i,j} = (m^t_{i,i-1}m^{t-1}_{i-1,i-2})h^{t-2}_{i-2,j} + (m^t_{i,i-1}m^{t-1}_{i-1,i-1} + m^t_{i,i}m^{t-1}_{i,i-1})h^{t-2}_{i-1,j} + m^t_{i,i}m^{t-1}_{i,i}h^{t-2}_{i,j} \tag{14}$$

Since the first column of the matrix $\mathbf{H}_t$ encodes the degree distribution, it suffices to calculate the values of $h^t_{i,1}$ in order to estimate $p_{k,t}$. In fact, $p_{k,t}$ (i.e., the $(k+1)^{\text{th}}$ entry of $\mathbf{H}_t$) is equal to $h^t_{k+1,1}$. In the following, we shall attempt to expand certain values of $h^t_{k+1,1}$ in order to detect the presence of a pattern (if any) in these values. In particular, let us investigate two cases of $h^1_{2,1}$ and $h^2_{2,1}$ from Fig. 7. We have

$$h^1_{2,1} = m^1_{2,1}h^0_{1,1} + m^1_{2,2}h^0_{2,1} = \left(1-\frac{1}{N}\right)\left(\frac{1}{\gamma+N}\right) + \left(\frac{N-1}{\gamma+N}\right)\left(\frac{1}{N}\right) \tag{15}$$

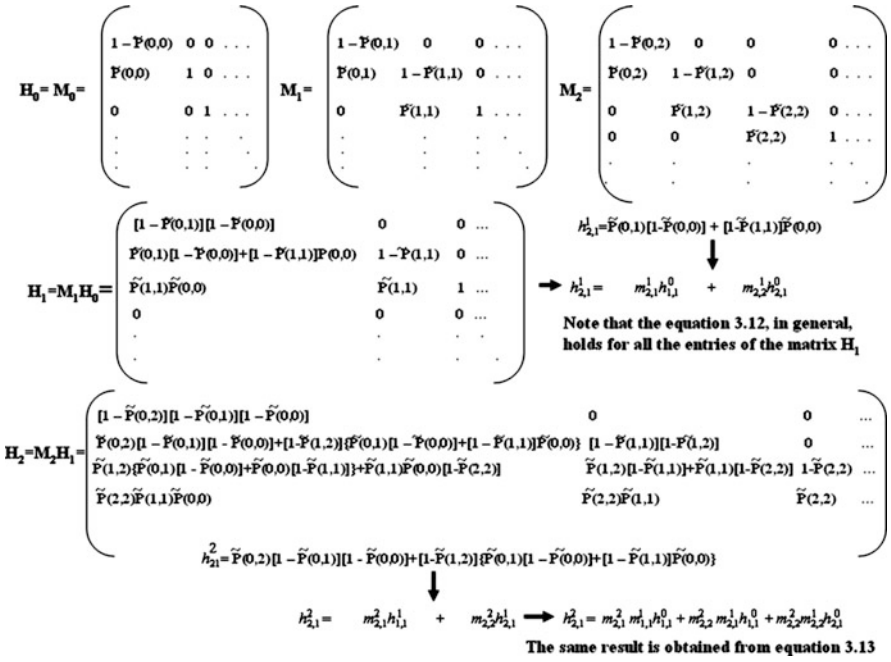

Fig. 7 A few steps showing the calculations of Eqs. 13 and 14

or,

$$h^1_{2,1} = 2\frac{(N-1)}{(\gamma+N)N} \tag{16}$$

Similarly,

$$h^1_{2,1} = m^2_{2,1}m^1_{1,1}h^0_{1,1} + m^2_{2,2}m^1_{2,1}h^0_{1,1} + m^2_{2,2}m^1_{2,2}h^0_{2,1} \tag{17}$$

or,

$$h^1_{2,1} = 3\frac{(\gamma+N-1)(N-1)}{(2\gamma+N)(\gamma+N)N} \tag{18}$$

A closer inspection of Eqs. 16 and 18 reveals that the pattern of evolution of this row, in general, can be expressed as

$$p_{k,t} = \binom{t}{k} \frac{\prod_{x=0}^{k-1}(\gamma x+1)\prod_{y=0}^{t-1-k}(N-1+\gamma y)}{\prod_{w=0}^{t-1}(\gamma w+N)} \tag{19}$$

for $0 \leq k \leq t$ and $p_{k,t} = 0$ otherwise. Further, we define the special case $\prod_{z=0}^{-1}(\ldots) = 1$. Note that if we now put $t = 2, k = 1$ and $t = 3, k = 1$ in (19) we recover Eqs. 16 and 18 respectively.

Equation 19 is the exact solution of the Eq. 6 for the initial condition $p_{k,t=0} = \delta_{k,0}$. Therefore, this is the analytical expression for the degree distribution of the consonant nodes in PlaNet$_{theo}$ for $\mu = 1$.

In the limit $\gamma \to 0$ (i.e. when the attachments are completely random) Eq. 19 takes the form

$$p_{k,t} = \binom{t}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{t-k} \tag{20}$$

for $0 \leq k \leq t$ and $p_{k,t} = 0$ otherwise.

On the other hand, when $\gamma \to \infty$ (i.e., when the attachments are completely preferential) the degree distribution of the consonant nodes reduces to

$$p_{k,t} = \left(1 - \frac{1}{N}\right) \delta_{k,0} + \frac{1}{N} \delta_{k,t} \tag{21}$$

Solution for $\mu > 1$

In the previous section, we have derived an analytical solution for the degree distribution of the consonant nodes in PlaNet$_{theo}$ specifically for $\mu = 1$. However, note that the value of μ is greater than 1 (approximately 21) for the real network (i.e., PlaNet). Therefore, one needs to analytically solve for the degree distribution for values of μ greater than 1 in order to match the results with the empirical data. Here we attempt to generalize the derivations of the earlier section for $\mu > 1$.

We assume that $\mu \ll N$ (which is true for PlaNet) and expect Eq. 6 to be a good approximation for the case of $\mu > 1$ after replacing $\widetilde{P}(k,t)$ by $\widehat{P}(k,t)$ where $\widehat{P}(k,t)$ is defined as

$$\widehat{P}(k,t) = \begin{cases} \frac{(\gamma k+1)\mu}{\mu\gamma t+N} & \text{for} \quad 0 \leq k \leq \mu t \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

The term μ appears in the denominator of the Eq. 22 for $0 \leq k \leq \mu t$ because, in this case the total degree of the consonant nodes in PlaNet$_{theo}$ at any point in time is μt rather than t as in Eq. 7. The numerator contains a μ since at each time step there are μ edges that are being incorporated into the network rather than a single edge.

The solution of Eq. 6 with the attachment kernel defined in Eq. 22 can be expressed as

$$p_{k,t} = \binom{t}{k} \frac{\prod_{x=0}^{k-1} (\gamma x + 1) \prod_{y=0}^{t-1-k} (\frac{N}{\mu} - 1 + \gamma y)}{\prod_{w=0}^{t-1} (\gamma w + \frac{N}{\mu})} \tag{23}$$

for $0 \leq k \leq \mu t$ and $p_{k,t} = 0$ otherwise.

Given that $\mu \ll N$ we can neglect the term containing μ/N in the Eq. 23 and express the rest using factorials as

$$p_{k,t} = \frac{t!\eta!(t-k+\eta-\gamma^{-1})!(k-1+\gamma^{-1})!\gamma^{-1}}{(t-k)!k!(t+\eta)!(\eta-\gamma^{-1})!(\gamma^{-1})!} \tag{24}$$

where $\eta = N/\mu\gamma$. Approximating the factorials using Stirling's formula (see [1] for a reference), we get

$$p_{k,t} = \widetilde{A}(t,\gamma,\eta)\frac{(k-1+\gamma^{-1})^{k-1+\gamma^{-1}+0.5}(t-k+\eta-\gamma^{-1})^{t-k+\eta-\gamma^{-1}+0.5}}{k^{k+0.5}(t-k)^{t-k+0.5}} \tag{25}$$

where

$$\widetilde{A}(t,\gamma,\eta) = \frac{t^{t+0.5}\eta^{\eta+0.5}\gamma^{\gamma^{-1}-0.5}e}{\sqrt{2\pi}(t+\eta)^{t+\eta+0.5}(\eta-\gamma^{-1})^{\eta-\gamma^{-1}+0.5}} \tag{26}$$

is a term independent of k.

Since we are interested in the asymptotic behavior of the network such that t is very large, we may assume that $t \gg k \gg \eta > \gamma^{-1}$. Under this assumption, we can re-write the Eq. 25 in terms of the fraction k/t and this immediately reveals that the expression is approximately a β-distribution in k/t. More specifically, we have

$$p_{k,t} \approx \widehat{A}(t,\eta,\gamma)\mathrm{B}(k/t;\gamma^{-1},\eta-\gamma^{-1}) = \widehat{A}(t,\eta,\gamma)(k/t)^{\gamma^{-1}-1}(1-k/t)^{\eta-\gamma^{-1}-1} \tag{27}$$

where $\mathrm{B}(z;\alpha,\beta)$ refers to a β-distribution over variable z. We can generate different distributions by varying the value of γ in Eq. 27. We can further compute $P_{k,t}$ (i.e. the cumulative degree distribution) using Eqs. 1 and 27 together.

References

1. Abramowitz, M., & Stegun, I. (Eds.). (1974). *Handbook of mathematical functions*. New York: Dover Publications.
2. Albert, R., & Barabási, A. L. (2000). Topology of evolving networks: Local events and universality. *Physical Review Letters, 85*, 5234–5237.
3. Amaral, L. A. N., Scala, A., Barthélémy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences, 97*, 11149–11152.
4. Anderson, H. (1989). Markedness theory: The first 150 years. In O. M. Tomic (Ed.), *Markedness in synchrony and diachrony* (pp. 11–46). Berlin: Mouton de Gruyter.
5. Balter, M. (2003). Early date for the birth of indo-european languages. *Science, 302*(5650), 1490.
6. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.
7. Barabási., A. L., Jeong, H., Ravasz, R., Néda, Z., Vicsek, T., & Schubert, A. (2002). On the topology of the scientific collaboration networks. *Physica A, 311*, 590–614.
8. Blevins, J. (2004). *Evolutionary phonology: the emergence of sound patterns*. Cambridge/ New York: Cambridge University Press.

9. Boersma, P. (1998). *Functional phonology*. The Hague: Holland Academic Graphics.
10. Bulmer, M. G. (1979). *Principles of statistics*. New York: Dover Publications.
11. Bybee, J. L. (1995). Diachronic and typological properties of morphology and their implications for representation. In L. B. Feldman (Ed.), *Morphological aspects of language processing* (pp. 225–246). Hillsdale: Lawrence Erlbaum Associates.
12. Caldarelli, G., & Catanzaro, M. (2004). The corporate boards networks. *Physica A, 338*, 98–106.
13. Choudhury, M., and Mukherjee, A. (2009). The Structure and Dynamics of Linguistic Networks. In Dynamics on and of Complex Networks: Applications to Biology, Computer Science, Economics, and the Social Sciences, Ganguly, N., Deutsch, A., and Mukherjee, A., (eds.), Birkhauser, Springer, Boston, 145–166, ISBN: 978-0-8176-4750-6.
14. Clements, G. N. (2008). The role of features in speech sound inventories. In E. Raimy & C. Cairns (Eds.), *Contemporary views on architecture and representations in phonological theory*. Cambridge, MA: MIT Press.
15. de Boer, B. (1999). *Self-organisation in vowel systems*. Ph.D. thesis, AI Lab, Vrije Universiteit Brussel.
16. Dorogovtsev, S. N., & Mendes, J. F. F. (2003). *Evolution of networks: From biological nets to the internet and WWW*. Oxford: Oxford University Press.
17. Eubank, S., Guclu, H., Kumar, V. S. A., Marate, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature, 429*, 180–184.
18. Evans, T. S., & Plato, A. D. K. (2007). Exact solution for the time evolution of network rewiring models. *Physical Review E, 75*, 056101.
19. Ferrer-i-Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences, 268*(1482), 2261–2265.
20. Guillaume, J. L., & Latapy, M. (2004). Bipartite structure of all complex networks. *Information Processing Letters, 90*(5), 215–221.
21. Hayes, B. (2004). Phonological acquisition in optimality theory: The early stages. In R. Kager, W. Zonneveld, & J. Pater (Eds.), *Fixing priorities: Constraints in phonological acquisition*. Cambridge: Cambridge University Press.
22. Hinskens, F., & Weijer, J. (2003). Patterns of segmental modification in consonant inventories: A cross-linguistic study. *Linguistics, 41*(6), 1041–1084.
23. Hughes, B. D. (1995). *Random walks and random environments: Random walks* (Vol. 1). New York: Oxford Science Publications.
24. Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Optimization models of sound systems using genetic algorithms. *Computational Linguistics, 29*(1), 1–18.
25. Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.
26. Lambiotte, R., & Ausloos, M. (2005). N-body decomposition of bipartite networks. *Physical Review E, 72*, 066117.
27. Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language, 48*, 839–862.
28. Lindblom, B. (1986). Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology* (pp. 13–44). Orlando, FL: Academic.
29. Lindblom, B., & Maddieson, I. (1988). Phonetic universals in consonant systems. In M. Hyman & C. N. Li (Eds.), *Language, speech, and mind* (pp. 62–78). London/New York: Routledge
30. Maddieson, I. (1980). *Working Papers in Phonetics* (UCLA N050). Los Angeles: Department of Linguistics, University of California
31. Maddieson, I. (1984). *Patterns of sounds*. Cambridge/New York: Cambridge University Press.
32. Maddieson, I. (1999). In search of universals. In *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 2521–2528). Berkley: University of California
33. Mukherjee, A., Choudhury, M., Basu, A., and Ganguly, N. (2008). Modeling the Structure and Dynamics of the Consonant Inventories: A Complex Network Approach , In the proceedings of COLING(08), 601–608, Manchester

34. Newman, M. E. J. (2001). Scientific collaboration networks. *Physical Review E, 64*, 016131.
35. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2), 167–256.
36. Ohkubo, J., Tanaka, K., & Horiguchi, T. (2005). Generation of complex bipartite graphs by using a preferential rewiring process. *Physical Review E, 72*, 036120.
37. Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford: Oxford University Press.
38. Peltomäki, M., & Alava, M. (2006). Correlations in bipartite collaboration networks. *Journal of Statistical Mechanics: Theory and Experiment, 1*, P01010.
39. Pericliev, V., & Valdés-Pérez, R. E. (2002). Differentiating 451 languages in terms of their segment inventories. *Studia Linguistica, 56*(1), 1–27.
40. Peruani, F., Choudhury, M., Mukherjee, A., & Ganguly, N. (2007). Emergence of a non-scaling degree distribution in bipartite networks: A numerical and analytical study. *Europhysics Letters, 79*(2), 28001.
41. Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Malden/Oxford: Wiley-Blackwell.
42. Ramasco, J. J., Dorogovtsev, S. N., & Pastor-Satorras, R. (2004). Self-organization of collaboration networks. *Physical Review E, 70*, 036106.
43. Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics, 25*, 255–286.
44. Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425–440.
45. Sneppen, K., Rosvall, M., Trusina, A., & Minnhagen, P. (2004). A simple model for self-organization of bipartite networks. *Europhysics Letters, 67*, 349–354.
46. Souma, W., Fujiwara, Y., & Aoyama, H. (2003). Complex networks and economics. *Physica A, 324*, 396–401.
47. Strogatz, S. H. (2001). Exploring complex networks. *Nature, 410*, 268–276.
48. Trubetzkoy, N. (1931). Die phonologischen systeme. *TCLP, 4*, 96–116.
49. Trubetzkoy, N. (1969). *Principles of phonology*. Berkeley: University of California Press.
50. Wang, W. S.-Y. (1971). The basis of speech. In C. E. Reed (Ed.), *The learning of language*. New York: Appleton-Century-Crofts
51. Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440–442.
52. Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Part II
Classifying Words and Mapping Them to Meanings

From Cues to Categories: A Computational Study of Children's Early Word Categorization

Fatemeh Torabi Asr, Afsaneh Fazly, and Zohreh Azimifar

Abstract Young children exhibit knowledge of abstract syntactic categories of words, such as noun and verb. A key research question is concerned with the type of information that children might use to form such categories. We use a computational model to provide insights into the (differential and cooperative) role of various information sources (namely, distributional, morphological, phonological, and semantic properties of words) in children's early word categorization. Specifically, we use an unsupervised incremental clustering algorithm to learn categories of words using different combinations of these information sources, and determine the role of each type of cue by evaluating the quality of the resulting categories. We conduct two types of experiments: First, we compare the categories learned by our model to a set of gold-standard part of speech (PoS) tags, such as verb and noun. Second, we perform an experiment which simulates a particular language task similar to what performed by children, as reported in a psycholinguistic study by Brown (J Abnor Soc Psychol 55(1):1–5, 1957). Our results suggest that different categories of words may be recognized by relying on different types of cues. The results also indicate the importance of knowledge of word meanings for their syntactic categorization, and vice versa: Addition of semantic information leads to the construction of categories

F.T. Asr (✉) · Z. Azimifar
Computer Science and Engineering Department, Shiraz University, Shiraz, Iran
e-mail: torabi@cse.shirazu.ac.ir; azimifar@cse.shirazu.ac.ir

A. Fazly
School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O.Box 19395-5746, Tehran, Iran
e-mail: afsaneh.fazly@gmail.com

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_4,

with a better match to the gold-standard parts of speech. On the other hand, our model (like children) can predict the semantic class of a word (e.g., action or object) by drawing on its learned knowledge of the word's syntactic category.

1 Introduction

Language acquisition is perhaps one of the most difficult tasks that children face early in their development. Children must process the auditory input in order to identify sentence and word boundaries, they need to map words onto their meanings, and must discover the syntactic constraints for combining individual words into meaningful phrases and sentences. There is abundant evidence that these processes of speech segmentation, word–meaning mapping, and syntax acquisition are intertwined. In particular, children are known to use their knowledge of the syntactic structure of utterances to help them identify words in speech (e.g., [13]), and also to narrow down possible meanings of words (e.g., [21]). In this study, we focus on the acquisition of syntactic categories (e.g., noun and verb) in young children, which have been argued to be a necessary first step for learning the structure of language. In fact, there is evidence that children have a good knowledge of many such abstract categories by 6 years of age (e.g., [16]).

Computational modeling has recently been accepted as a new paradigm for the study of various aspects of human language acquisition and processing. In particular, many computational models have been proposed to answer several important questions regarding the development of syntactic categories in young language learners: E.g., what is the nature of the learned categories; whether such categories are learnable from usage data; and what sort of cues in the input facilitate their acquisition. Many of these models, however, make assumptions about the underlying learning mechanisms or about what a learner can extract from the input, which are arguably beyond the abilities of young children. Other more recently-developed and cognitively-plausible models focus mainly on one particular type of cue for the acquisition of syntactic categories, overlooking other important sources of information. What is lacking is a comprehensive investigation of the interplay of the various types of cues on early syntactic category learning. The study reported here is a first attempt at addressing this gap using a computational modeling approach.

This article is an updated and extended version of our previous work (reported in [2]). The article is organized as follows: We introduce the existing psycholinguistic and computational studies focusing on syntactic category learning in Sect. 2, and then provide an overview of our proposed study in Sect. 3. Section 4 presents the components of the model, and Sect. 5 describes the experimental setup. Results are reported in two separate Sects. 6 and 7, the former including evaluations of the learned categories through comparison with parts of speech (PoS), and the latter examining the usefulness of the learned categories in a language task. Our discussions and analyses of the results are presented in the final concluding section.

2 Related Work

Psycholinguistic studies on early language acquisition have shown that children can recognize shared syntactic and/or semantic properties among words, and that they use this knowledge (through categorization and generalization) to learn new words [3, 4, 11, 16, 26]. Brown (1957) [4] and Berko [3] investigate how children's knowledge of syntax and semantics interact. In Brown's experiment, children are asked to predict the meaning of a novel (often made-up) word appearing in a familiar syntactic and/or morphological context. For example, children should predict what *dax* means in *This is a dax* or in *He is daxing*. Whereas Brown focuses on semantic predictions using syntax, Berko investigates how children make syntactic decisions on the basis of semantic information. For example, children are shown a picture of a cartoon animal while they hear *This is a wug*. When presented with a similar picture of two such cartoons, children are expected to fill in the empty spot in an utterance such as *There are two* $\cdots$ with a (morphologically) appropriate word such as *wugs*.

Results of the above seminal studies suggest that young children (3–7 years old) categorize words in the course of their language development, and that they derive general rules about entities within a category. Since then, many researchers have tried to better understand the acquisition of word categories in young children. Gelman and Taylor [11] investigate how 2-year-old children interpret novel nouns as proper or as common category names, depending on the syntactic context in which the noun is first introduced (e.g., *This is a Zav* or *This is Zav*), as well as on the type of the novel referent (e.g., animal-like or block-like toy). The results of this study show that these young children use both types of knowledge (linguistic context and real-world knowledge) to interpret new words. Samuelson and Smith [26] find that properties of objects, such as their shape or texture, affect children's assumptions about the syntactic categories of the names used to refer to the objects. For example, they find that children tend to use count nouns to refer to solid objects. Kemp et al. [16] study the developmental patterns of the determiner and adjective categories in 2–6-year-old children. Results of their experiments suggest that children's understanding of determiners and adjectives are built gradually over several years, possibly starting with constructions learned for individual lexical items.

An important question regarding children's early word categories is what cues facilitate their acquisition. To answer this question, two streams of research have emerged. One group of studies perform experiments on children or use large corpora of child-directed speech (CDS) to shed light on various aspects of category learning, and in particular on what types of cues are *available* in the learning environment of young children [12, 19, 20, 22]. We provide more details about these studies in Sect. 2.1. Another group of researchers have developed computational models to examine whether these cues are actually *useful* in learning syntactic categories, and whether categories built by a model resemble those learned by children [1, 5, 7, 8, 22, 23, 25, 27]. Section 2.2 explains these studies in more detail.

2.1 Experimental and Corpus-Based Studies

Several studies have looked into the availability and relevance of distributional word co-occurrence cues on the acquisition of early word categories. Gerken et al. [12] perform experiments on very young American (English-speaking) children who have had limited exposure to Russian. Children in this study show acceptable competence for distinguishing linguistic gender categories based on distributional information available in the (Russian) training data provided to them. These results suggest that, from relatively early in language development, distributional information is used by learners as a strong cue to word categories. Mintz [19] investigates the usefulness of *frequent frames*—co-occurrence patterns of words in sentences—in syntactic categorization: Words that appear in the same frequent frame are categorized together. Mintz's analysis of CDS data shows that even very simple three-word frames yield reasonably accurate adult-like syntactic categories.

Others have examined the type of phonological and morphological information found in CDS that are likely to be relevant to syntactic categories. Monaghan et al. [20] present a comprehensive investigation of phonological cues that mark syntactic categories in a number of languages. They show that simple phonological cues are available to young children (in CDS); and that they are particularly helpful for the identification of the syntactic category of low-frequency words. Through an extensive analysis of CDS in several languages (namely, English, Dutch, French, and Japanese), [22] show that simple approximations about the morphological structure of a word—i.e., the first and last phonemes of the word—significantly correlate with its syntactic category.

2.2 Related Computational Models

Computational modeling is considered as a powerful tool for the study of language acquisition, and hence many computational models have been proposed to study the *learnability* of syntactic categories from a usage-based point of view [15, 24]. Several existing computational models cluster words into syntactically (and semantically) similar categories, by drawing on distributional cues (i.e., word co-occurrence statistics) extracted from usages of words in context [8, 25, 27]. The results of these studies suggest that abstract word categories are learnable, and that distributional cues are a useful source of information for this purpose. The connectionist model of Onnis and Christiansen [22] can infer the lexical category of a novel word by drawing only on simple phonological/morphological information. The proposed model is not fully unsupervised, as it needs to be trained with an initial small sample of words labeled for lexical category. Nonetheless, the results of this study suggest that simple learning mechanisms combined with simple phonetic information (that are easily available to children) are useful in the acquisition of syntactic categories.

A few models have been proposed that are meant to be cognitively more plausible—that is, they are fully unsupervised, they process data incrementally, they categorize individual occurrences/usages of a word (as opposed to word types as in the models above), and they do not need to specify the number of categories beforehand. The incremental model of Cartwright and Brent [5] builds categories by finding common usage patterns across sentence-length templates. The proposed algorithm is very effective for discovering categories in artificial languages. However, the model uses the full sentence as a contextual unit, and hence is not sufficiently flexible to handle noise (which is prevalent in naturalistic child-directed data).

Several probabilistic models have been put forward to provide the flexibility needed for handling noise. Parisien et al. [23] propose a Bayesian clustering model which can handle ambiguity, and exhibits some of the developmental trends observed in children (e.g. the order of acquisition of different parts of speech). In order to overcome sensitivity to variability in context, they introduce a "bootstrapping" component—where the model uses its own learned knowledge of categories in future categorizations—as well as a periodical cluster reorganization component. These mechanisms improve the overall performance of the model, but also make it more complex. Alishahi and Chrupała [1] and Chrupała and Alishahi [7] propose incremental models of lexical category learning that can efficiently process naturalistic utterances, and that can over time build robust categories from little usage data.

3 Overview of This Study

The existing computational studies (presented in the previous section) demonstrate that syntactic categories can be learned from naturalistic word usages, by drawing on the kinds of information that children are known to be sensitive to. Nonetheless, these studies have shortcomings that need to be addressed. Some of the proposed models incorporate batch learning mechanisms, such as hierarchical clustering, which are not intended to be cognitively plausible (e.g., [8, 25, 27]). In addition, these models partition the vocabulary (word types) into a set of non-overlapping clusters, and hence do not account for words that belong to more than one category. Moreover, most existing studies have focused on the usefulness of one type of cue, mainly distributional (as in [1, 5, 7, 8, 23, 25, 27]), and rarely on another type of cue (e.g., phonological, as in [22]).

We present a computational study that investigates the role of different types of language-internal (namely, distributional, morphological, phonological) and language-external (i.e., semantic) cues in the acquisition of syntactic categories. Specifically, we apply a simple incremental clustering algorithm (a slightly modified version of the model proposed by Alishahi and Chrupała [1]) on naturally-occurring English child-directed utterances to study the effect of each of the above-mentioned

cue types on category learning. We choose the above-mentioned categorization algorithm because it incrementally categorizes word *usages*; and moreover, the model relies on simple similarities among word properties, and can easily incorporate any types of cues to build the categories. Our goal is to provide a detailed analysis of the interplay of multiple cue types in the acquisition of early syntactic categories. We thus compare categories provided by different combinations of cue types. In addition, we examine whether there are any meaningful correlations between each cue type and the acquisition of each syntactic category.

To evaluate the learned categories, we perform two types of analyses: First, we compare the categories induced by our model to a small set of gold-standard PoS tags (e.g., noun and verb). This is not an ideal but a common evaluation of the similarity between the knowledge acquired by a computational model and the formal word categories in adult grammar. Second, we present a set of semantic prediction experiments where we examine the usefulness of the categories for predicting the meanings of novel words. In one such experiment, we also compare the behaviour of our model to that of children as reported in the psycholinguistics literature. Results of our experiments reveal the effect of different cues in recognizing words from different syntactic categories: Morphological and phonological properties of words, as well as their semantic properties, help increase the overall categorization performance of the model when used in combination with contextual (word co-occurrence) features. (We thus take our baseline performance to be the model's performance when only co-occurrence features are used.) Our results also show that morphological and phonological features help identify verbs, whereas nouns are better categorized when these features are ignored. Our results also reveal interesting syntax–semantics interactions in the learned categories: word meanings prove to be useful for syntactic category learning; and, moreover, the syntactic context (as well as the morpho-/phonological properties) of a novel word are shown to be helpful for predicting its semantics.

Best performance of the model in recognizing syntactic categories is obtained when semantic features are taken into account that in turn shows a strong syntactic–semantic coherence of the constructed categories. Another evidence for this claim is the result obtained in our semantic prediction task. Although the model in this experiment is trained solely on syntactic features, its performance in predicting semantically similar clusters for new words is much above the chance level. Furthermore, in the simulation of Brown's experiment, the model's behavior has a good match with that of children in distinguishing objects from actions (when only limited syntactic information is provided to them).

To summarize, we study the effect of several different types of (language-internal and language-external) cues on early word categorization. To do so, we use an existing incremental similarity-based clustering algorithm that can easily incorporate different combinations of cues to form categories of words. We incorporate only very simple cues that are known to be both *available* in the input that children receive, and *accessible* by very young children (i.e., easy to extract from their learning environment). Following previous studies, we evaluate the resulting

clusters in each experimental condition by comparing them to a set of gold-standard part of speech tags. In addition, we also simulate several language tasks for which young children need to use their knowledge of categories.

4 Components of the Categorization Model

We first present the categorization algorithm of Alishahi and Chrupała [1] that we slightly modify to fit our purpose (Sect. 4.1); and then explain how we extract the different types of cues from naturally-occurring child-directed utterances to be used for learning word categories (Sect. 4.2).

4.1 Categorization Algorithm

The unsupervised clustering algorithm proposed by Alishahi and Chrupała [1] works based on contextual similarities among words.The algorithm is incremental in the sense that it processes words one by one, discarding each word after clustering. For each newly-observed *frame* (a target *head-word* along with its left and right neighboring words), if the similarity to all of the already-shaped clusters is less than a predefined threshold, a new cluster is formed. Otherwise, the target word is assigned to the most similar cluster.

We choose this algorithm for our study since it is unsupervised, incremental, flexible, and simple. We modify the algorithm in two ways: (1) the original algorithm of Alishahi and Chrupała [1] includes a phase in which clusters are merged if they are sufficiently similar. To keep the algorithm simple, we remove this step and adjust the final number of clusters by the help of a similarity threshold; (2) our frames are composed of different types of *features*; we thus need to modify the calculation of the similarity score in order to accommodate for more than one set of features. We calculate the similarity between a frame f (consisting of a variety of features) and a cluster C (a group of frames) as:

$$Sim(\mathrm{f}, C) = \sum_{i \in \mathcal{F}} \omega_i * Sim_i(\mathrm{f}, C) \tag{1}$$

in which $\mathcal{F}$ is the set of all features considered as part of our frames, $Sim_i(\mathrm{f}, C)$ shows the similarity of frame f to cluster C with respect to the ith feature, and the weight ω_i determines the relative contribution of the ith feature in calculating the overall similarity between a frame and a cluster. Weights for all features need to sum to 1, i.e., $\sum_i \omega_i = 1$. Details of the modified algorithm are given in Algorithm 1.

Algorithm 1 Incremental word clustering

initialize set of clusters $\mathcal{K} = \emptyset$
for every frame f **do**
  $C_M = argmax_{C \in \mathcal{K}} Sim(\text{f}, C)$
  if $Sim(\text{f}, C_M) \geq \theta$ **then**
    Add frame f to cluster C_M
  else
    Construct a new cluster for frame f
  end if
end for

(This algorithm is a modification of the one proposed by Alishahi and Chrupała [1]).

4.2 Cues Used in Categorization

As previously mentioned, children are known to group words into categories by drawing on a number of different types of cues. In this study, we include four different sources of information:

Distributional information about word co-occurrences: This kind of information has been reported to be reliable and very important in syntactic categorization [1, 8, 19, 23, 25, 27]. Some of these studies report that words closer to a word are more informative about its category. Therefore, we take one word from each side of a target head-word as its co-occurrence features. For example, considering the following sentences hints the model to group *cat* and *table* together since they share similar co-occurrence features:

There is a *cat* in the basket
We need a *table* in our kitchen
The *cat* is in the house
The *table* is in good condition

In our framework, each co-occurring word is considered as an independent feature when determining similarity between a word (frame) and a cluster (as done in many previous studies, and in contrast to representations such as "frequent frames" of Mintz [19]). For example, even if the two tokens *cat* and *table* did not share the preposition *in*, they should still be considered as somewhat similar because of the preceding determiner *a* that they have in common.

Phonological information: Words belonging to the same syntactic category tend to have common phonological properties. For example, by analyzing the child-directed utterances, Monaghan et al. [20] show that verbs and nouns differ with respect to several phonological features, including the number of syllables. The study done by Monaghan et al. [20] focuses on the relevance of syntactic categories and a large number of word-level, syllable-level, and phoneme-level phonological properties. Here, we only focus on two of the simplest word-level phonological properties that are assumed to be readily accessible by young children, namely the

length of a word in terms of the numbers of syllables and phonemes (number of letters are taken to approximate the number of phonemes in a word).

Morphological information: It has been shown that English affixes, such as *-ing* in verbs, can provide strong clues to the identification of syntactic categories, and that such information is abundant in child-directed speech [22]. Nonetheless, it is not clear whether one can assume that children have access to such accurate morphological knowledge about words and categories prior to syntactic category learning. Inspired by the work of Onnis and Christiansen [22], we use the last phoneme (ending) of the words as an approximation to the morphological affixes.[1]

Semantic properties: In addition to the previously mentioned information that are directly available in the utterances, children might use some notion of word meanings when finding similarities among words. We use a set of semantic features previously used by Fazly et al. [9] in a computational model of word learning. More details about this resource are provided in Sect. 5.2.

In our experiments, we use different combinations of the above-mentioned features (cues). Head word, co-occurrence properties, morphology and phonology are considered as language-internal features, whereas semantic features are taken as language-external properties of the words. Further details about the extraction of these features is presented in the following section (Sect. 5.2).

5 Experimental Setup

We first provide information on the corpus we use in our experiments (Sect. 5.1), and then explain how we extract our features from the utterances (Sect. 5.2). Finally we present some details about setting the parameter values in the model (Sect. 5.3).

5.1 Corpus

We use as our input corpus naturally-occurring utterances similar to what children receive. Specifically, our input data (both for training and test) is obtained from the Manchester corpus [28], one of the English subsets in the CHILDES database [18]. This corpus contains conversations between parents/caregivers and 12 British children between the ages of 1;8 (years;months) and 3;0.[2] For training, we randomly

[1] In earlier experiments, we also included the first phoneme (beginning) of a word—a feature also considered by Onnis and Christiansen [22]. In our initial evaluations, we found that the inclusion of this feature did not affect the performance, and hence excluded it from further consideration.

[2] Authors are grateful to Christopher Parisien for providing them with a preprocessed version of this corpus.

Head:	*ball*
Cooc:	*a*, *in*
Phon:	1 syllable, 4 letters
Morph:	/*l*/
Sem:	{ GAME EQUIPMENT#1, EQUIPMENT#1, INSTRUMENTALITY#3, ARTIFACT#1, WHOLE#2, OBJECT#1, PHYSICAL ENTITY#1, ENTITY#1, BALL#1 }

Fig. 1 Sample frame extracted for head word *ball* from the utterance "There is a *ball* in the basket"

choose a number of child-directed utterances from the conversations of all 12 children such that the chronological order of the utterances is maintained and the utterances contain only words selected from a limited vocabulary of 500 word types. When selecting this vocabulary, we ensure that their distribution in the corpus matches the Zipfian distribution, so that our results are not biased towards words from certain frequency ranges. We extract 50,000 *frames*—each containing a target word to be categorized as well as some features—from these filtered utterances, and use them as our training set. We limit the size of the vocabulary because some feature values must be determined manually. Moreover, in one experimental task, we need access to natural novel words not previously seen in the training corpus, instead of artificially made-up words used in many psychological experiments. As our test data, we thus select 2,000 frames such that the target words to be categorized are novel for the model (i.e., not in the vocabulary of 500 words).

5.2 Feature Extraction

From each utterance (in either the training or test data), we extract a number of frames to be clustered. As previously explained, each frame contains a head word (the target word to be categorized), plus other features including: two co-occurrence (Cooc), two phonological (Phon), one morphological (Morph), and a set of semantic (Sem) features. A sample frame is shown in Fig. 1. The head word and the Cooc features can be directly extracted from the utterance. If any of the Cooc features is missing (i.e., when the target word is the first or the last word of the utterance), that feature value is set to "Null".[3]

For Phon and Morph features a phonemic representation of words as well as other phonological features are required. We extract two of these features (the ending phoneme, and the number of syllables) from the MRC Psycholinguistic Database, a publicly available resource built for the purpose of studies on child language [29].[4] If a word is not found in MRC, we set the values of the above features manually.

[3] The "Null" value is treated as a missing value for a feature.

[4] http://www.psych.rl.ac.uk/

For the third feature, namely the number of phonemes in a word, the number of letters is considered as an approximation.

Semantic features are extracted from the lexicon prepared by Fazly et al. [9]. In order to extract the right semantic features of a word usage, we first look at its PoS tag in the selected utterance and then look for the corresponding entry in the lexicon to retrieve its semantic feature set. Fazly et al. [9] have prepared their semantic lexicon using three different resources: For nouns and verbs, they take all the hypernyms for the first sense of the word in WordNet [10], where each hypernym is a set of synonym words (synset) tagged with their sense number. The first word in the synset of each hypernym is taken as a semantic feature for the target word. Fazly et al. augment the semantic features of verbs with extra properties taken from VerbNet [17]. For adjectives and closed-class words, the authors use semantic features provided by Harm [14].(See Fig. 1 for a sample set of semantic features associated with the word *ball*.)

5.3 Model Parameters

The model we use contains two sets of parameters: the weights ω_i in (1) used to determine the relative contribution of features in measuring similarities, and a similarity threshold θ used to decide whether to create a new cluster for a given frame. We set the weights ω_i uniformly, assigning equal weights to all features. Clearly the value of θ affects the population of generated clusters: a low value increases the likelihood of grouping more words, hence decreasing the total number of clusters. We assign different values to this parameter in various experimental conditions (i.e., different combinations of features), such that we maintain the total number of clusters generated in each condition within a desired range.

We take into account two different ways of measuring $Sim_i(\mathrm{f}, C)$ in (1) depending on feature i. For categorical features (Head, Cooc, Morph and Sem) we use the cosine similarity of the feature vectors, which is widely used in similar clustering algorithms. A vector representing a categorical feature such as Head is of the size of word types in the corpus. For a sample frame f the vector includes 0 in all elements except where the value of Head in that frame is presented. For our numerical feature (Phon) we use the Euclidean distance.

6 Discovering Syntactic Categories

This section explains the first set of our experiments which analyze the role of different language-internal and language-external cues in syntactic categorization. First, a detailed description of our evaluation strategy is given. Then, we present a discussion of the performance of our model in acquiring categories similar to gold-standard parts of speech when different combinations of the cues/features are used.

6.1 Evaluation Strategy

To examine the contribution of each cue type in syntactic categorization, we evaluate the quality of the clusters built by using different combinations of features in a *novel word categorization* task. Specifically, we train our model (on the training corpus) in five different conditions, i.e., using one of the following feature combinations: Cooc, Cooc + Morph, Cooc + Phon, Cooc + Sem, and finally Cooc + Morph + Phon + Sem (Note that the Head feature is always part of the frame to be categorized; we omit it from the title of the conditions for the ease of exposition only). We then determine the effect of each feature set by examining the performance of the model to infer the category of a number of novel (previously unseen) test words.

After the training phase, each learned cluster is labeled with a part of speech tag as follows. Words in the Manchester corpus are tagged with a fine-grained set of parts of speech. We convert these to a coarser grained version (also used by Parisien et al. [23]), including 11 tags, namely, Noun, Verb, Adjective, Adverb, Determiner, Negation, Infinitive, Auxiliary, Conjunction, Preposition, and Others. (Note that the resulting categories do not necessarily need to match the conventional adult-like categories put forth by linguists. Nonetheless, as a first-line evaluation, we compare the categories learned by our model to this gold-standard set of parts of speech.) Each cluster label is therefore assigned based on the majority label among all its members. For example, a cluster containing 30 nouns, 90 verbs, and 20 adjectives is labeled as a Verb cluster.

During the test phase the model does not create new clusters, but assigns each novel test word to one of the clusters (the one that is most similar to it) formed in the training phase. The test word is then assigned the same part-of-speech label as the selected cluster, and this label is compared with the 'true' syntactic category of the word that is the gold-standard tag associated with it in the corpus. We report *accuracy* measured as the proportion of test words assigned to their correct category. We also look into the accuracy for different groups of words, such as verbs and nouns, as well as open-class (content) and closed-class (function) words.

Accuracy corresponds to token-based *precision* on test data, a widely used measure for evaluating clustering performance along with *recall*. Precision is a good indicator of the homogeneity of the items within a cluster. We do not report recall values which are not informative in our experiments since the number of target labels (11 PoS tags) is very small in comparison to the number of resulting clusters. In fact, precision values show the significant differences in the performance of the model over different experiments while recall values are very small for all experiments due to the large number of clusters in general.

6.2 Novel Word Categorization

This section presents our evaluation of the trained model in predicting the syntactic category of the 2,000 novel test words, using different combinations of features. Since distributional information has been reported to play a key role in determining

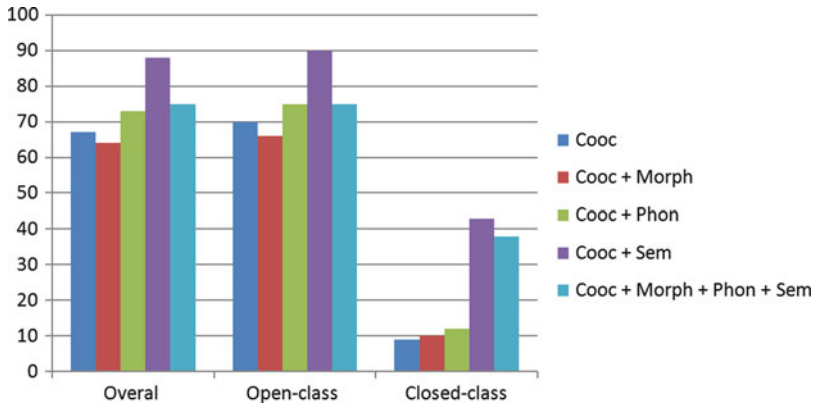

Fig. 2 %Accuracy of novel word categorization over five different combinations of features; Accuracy is reported separately for open-class and closed-class words

syntactic categories, the condition in which we use only Cooc features is considered as our baseline. We thus investigate the effect of the other features in cooperation with (or in comparison to) the Cooc features.

Note that the Head feature is not very useful for prediction (since test words are not seen during the training), and hence the model has to utilize other sources of information to determine the category of a word. Recall that we examine the performance of our model in category prediction using five different combinations of features. For each combination of features, Fig. 2 depicts the performance of the model on all test words, as well as separately for open-class and closed-class words.

Comparing the accuracy of the categorization model across different conditions is fair and meaningful only if the number of clusters is relatively close in all conditions. Generally speaking, allowing a large number of clusters makes the categorization more conservative (i.e., forming too many small clusters each containing one or a few word types that are very similar). Hence, in the training phase we assign different values to the similarity threshold θ to obtain approximately identical number of final clusters for each of the above-mentioned five conditions, e.g., between 250–300 clusters. This approach allows us to focus on the effect of different features involved in categorization while other factors are maintained constant across the experiments.[5]

Comparing the overall accuracies of the first four conditions suggests that employing phonological features improves the prediction of a novel word category while the morphological feature is not effective. In addition, Sem features are

[5]We have performed similar experiments with different ranges of cluster numbers, and found that the general patterns in results are similar. In Appendix A, we report the result of experiments in which we set the number of clusters within the range 346–500 (<500). In general we prefer fewer clusters (fewer than our vocabulary size) to allow for generalization. We expect the generalization ability of the model with 247–288 (<300) clusters to be reasonably good since more than 55 % of these clusters contain three or more word types in all conditions.

observed to have the most helpful contribution to the overall prediction accuracy. (We should note that our selection of semantic features is one among a variety of possible choices. Therefore, their effect in the categorization should be analyzed in a more conservative manner than for better defined features, i.e., language-internal properties). Best performance of the model is acquired by considering co-occurrence and semantic features, even better than the condition in which all features are used. This in turn shows that using a combination of several acceptable features do not necessarily result in a better accuracy than when applying them individually. Indeed, these results suggest that children may not always use all the cues in combination, but rather use each type of cue to identify words from a particular category. We provide further details on this prediction in the rest of the section.

The accuracy of category prediction for different classes of words are shown to be affected in different ways. The accuracy for closed-class words is very low, because such words neither share distinguishable word-internal (Phon and Morph) features nor are easily classified based on co-occurrence information. The latter is caused by the large-size vocabulary of open-class words appearing as neighbors of closed-class words. A well-known category of closed-class words are determiners whose frequency is very high in naturally occurring utterances while they have a rather small vocabulary size (e.g., *a, an, the*). Since nouns form a very diverse group and often appear to the right of determiners, they are too sparse to help group determiners into a category. On the other hand, determiners are good indicators of nouns and hence, as a co-occurrence feature they direct the model to choose the right category for a novel noun. Therefore, in general, closed-class words function as good co-occurrence features for open-class words while the reverse is not the case.

In order to better understand the performance of different cue types, the categorization accuracy for verbs, nouns, adjectives and determiners are separately reported in Fig. 3. In a quick look, we find the latter two categories more difficult for the model to learn. The lower accuracy of detecting novel determiners and adjectives in comparison to nouns and verbs suggests that the model can hardly employ its acquired knowledge to detect new instances of these categories. This, in fact, has been reported in psycholinguistic investigations of language acquisition. Children are known to acquire nouns and verbs easier and faster than other categories [16]. Above, we explained why determiners (as a typical example of closed-class words) are hard to be categorized by our model, but why are adjectives hard to learn as a group? Despite having distinctive word-internal features (e.g., ending *-y* as a morphological cue) the low frequency of adjectives in CDS might provide an explanation.[6] In addition, adjectives seem to share some distributional features with nouns (e.g., like nouns, adjectives may also be preceded by certain determiners).[7]

Interestingly, using Cooc features alone results in a better detection of novel nouns, whereas for verbs other types of information (i.e., Morph and Phon) are

[6]In both the training and test data less than 6 % of the vocabulary are adjectives.

[7]Note that although the results show that by using semantic features the prediction accuracies for adjectives and determiners are substantially improved, this effect is due to the nature of the semantic features for these words (taken from Harm [14]) and should be interpreted with caution.

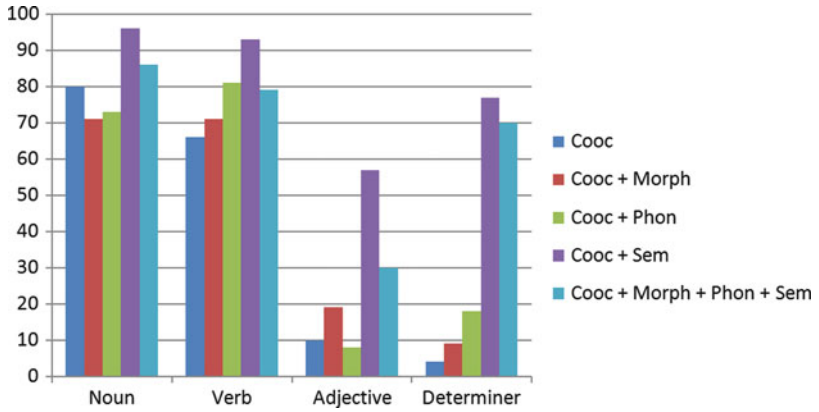

Fig. 3 %Accuracy of novel word categorization over five different combinations of features; Accuracy reported separately for four parts of speech

more helpful. Hence, even among open-class words, discovering different categories seem to rely on different types of information. This finding is also supported by the observation that, typically, context words such as determiners mark the appearance of nouns; and in contrast, verbs particularly share morphological and phonological properties (many verb forms end in *-ed* or *-ing*).[8]

In all of the experiments semantic features seem to play an effective role. These results direct our attention to the hypothesis that children use language-external information about words to induce lexical categories. In the following section we design experiments to further investigate the effect of syntax on semantic prediction—that is, to see how our model might use its learned (syntactic) categories to predict the meaning of a novel word.

7 Word Categorization and Semantic Prediction

One important task for examining the effectiveness of a word categorization model is that of *semantic prediction*—that is, predicting the semantics of a novel word appearing in a familiar syntactic context. Semantic prediction has been used both to test the ability of young children to form word categories [9], and to evaluate computational categorization models [7]. Our second set of experiments is thus designed to show how the categories learned by our model can help predict the semantics of a novel word. We first describe the results of an experiment in which we use the categories learned by using a combination of features to predict the semantics of the 2,000 novel words in our test data (Sect. 7.1). We then present a simulation of the Brown's [4] experiment (Sect. 7.2).

[8]Results of the novel word categorization experiment are included in the Appendix with more details.

7.1 Semantic Feature Prediction

We use the model trained on 50,000 input frames (using different combinations of features, as explained in Sect. 6) to predict the meaning of our 2,000 novel test words. For this experiment as well as the simulation of Brown's experiment (the following section), we build a *semantic profile* for each cluster formed during training. Specifically, at the end of the training phase each cluster will have a vector of semantic features which is an average over the semantic vectors of all members assigned to it. In the test phase of this experiment, we first assign each novel test frame to the most similar cluster—henceforth referred to as the *chosen* cluster—by comparing all the non-semantic features also used during training (i.e., a subset of Cooc, Morph, and Phon). We then measure the *semantic fit* between the target frame and the chosen cluster using two different measures: (1) the Mean Reciprocal Rank (MRR), which is a measure widely used to evaluate an information retrieval system; and (2) a novel measure we introduce, called Average Semantic Recall (ASR). We will now explain each of these measures in more detail. To measure semantic fit using either MRR or ASR, we rank all the learned clusters according to their semantic similarity to the target frame. Both MRR and ASR examine the rank of the chosen cluster in this ranked list, and assign a score to it that reflects how high in this list is the chosen cluster. MRR and ASR assign their highest score to a situation in which the semantically most similar cluster to the target frame (top cluster in the ranked list) is the chosen cluster (i.e., the syntactically and/or phonologically and/or morphologically most similar cluster).

Formally, MRR is calculated as follows:

$$\text{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \tag{2}$$

where n is the number of test words (here, 2,000), and r_i is the rank of the chosen cluster for the ith test word. Note that MRR does not take the total number of clusters in the ranked list into account, and hence may not be easily comparable across different experimental conditions in which the number of clusters varies. We propose a novel measure, which we call Average Semantic Recall (ASR), in order to alleviate this problem. MRR is designed for information retrieval, where the assumption is that users mostly notice the items at the very top of a ranked list. Hence the value of MRR decreases exponentially as the chosen cluster moves down the ranked list (due to averaging over the inverse ranks). In contrast, ASR averages over the actual ranks, and so its value decreases in a linear fashion, as in:

$$\text{ASR} = 1 - \frac{\sum_{i=1}^{n}(r_i - 1)}{n(k-1)} \tag{3}$$

where k is the total number of clusters learned by the model. MRR and ASR values range between 0 and 1, with 1 reflecting a really strong interaction between the

Table 1 Performance of the categorization model in predicting semantic features of 2,000 novel test frames; For the ease of exposition, we present multiples of ASR and MRR by 100

	ASR × 100	MRR × 100
Cooc	87	14
Cooc + Morph	82	11
Cooc + Phon	88	14
Cooc + Morph + Phon	81	11

syntax and semantics of the learned clusters (at least for the ones to which a test frame is assigned). Table 1 gives the MRR and ASR values in different conditions (in each the model is trained using a different combination of features).
As noted before, a similar experiment has been performed by Chrupała and Alishahi [7] but considering co-occurrence features only. Our results here confirm their findings: that syntactic context of a word (reflected in the Cooc features) closely interacts with the word's meaning. Our results, however, do not show any interactions between the morphological/phonological properties of words and their semantic features, independently of the syntactic context. Future research will need to investigate this, e.g., by considering richer morphological and phonological features.

7.2 Simulation of the Brown Experiment

In his seminal work, Brown [4] investigated children's ability to predict the meaning of a novel word by drawing on its syntactic context. The experiment is designed to examine how children distinguish three classes of words—namely, words referring to *Actions*, *Objects* or *Substances*—based on the syntactic structure of the sentences the words appear in. An examiner shows a picture to a child and utters a sentence to describe the picture. The picture displays an activity such as pouring a confetti-like material in a container. In other words, this picture depicts three major elements: an action (e.g., pouring), an object (e.g., a container), and a substance (e.g., the confetti). The sentence uttered by the examiner is designed to intentionally include one novel (unfamiliar) word in co-occurrence with several familiar words. In our example the examiner might say "You can see some *sib* in the photo", where *sib* is an unfamiliar word for the child. The child is then asked to choose "*some sib*" among three other pictures, each containing only one of the three elements previously displayed within the first picture. If the child selects the right picture (here, the one showing a substance similar to the confetti presented in the first picture) it is counted as a correct answer.

Based on the observations from this experiment, Brown [4] concludes that children tend to predict semantic properties of the novel words by taking advantage of syntax. In other words, frames such as "some *x*" work as a cue for the child to

Table 2 Performance of the categorization model in assigning different types of frames to three semantic categories

	Object	Substance	Action
Countable nouns	**45**	4	1
Uncountable nouns	41	**7**	2
Verbs	0	0	**50**

spell the category of the word x; therefore, semantic properties of x (e.g., relating to a substance) can be predicted by using its recognized category. In order to adapt the study performed by Brown [4] into our framework we first train the categorization model with 50,000 input frames based on Head, Cooc and Sem features (see Sect. 5 for details). In the test phase a number of frames including novel words as their Head features are presented to the model. Each of these test frames are selected from child-directed utterances such that its Head is a verb, a countable or an uncountable noun. We then build a semantic profile for each test frame, as a weighted average of the semantic profiles of the learned clusters, as in:

$$Sem(\mathrm{f}) = \sum_{C_i \in Clusters} Sim(\mathrm{f}, C_i) * Sem(C_i) \tag{4}$$

where $Sem(C)$ indicates the semantic profile of cluster C, weighted by $Sim(\mathrm{f}, C)$ which is calculated as in Eq. (1) on Page 87. (To measure $Sim(\mathrm{f}, C)$ we use only the Cooc features since they proved to be the most effective in our previous experiments.) Calculating the semantic profile of a novel word (frame) is thus affected by the relative similarity of the frame to the learned clusters: The more similar a cluster is to the frame, the more important role it plays in determining the semantics of the frame.

This semantic prediction is intended to simulate a child's (informed) guess about the semantics of a novel word in the experiment of Brown. As we stated, the child in the Brown's experiment is then exposed to three pictures each presenting only one element of the first picture. Analogously, in our model, after performing the above calculations, the obtained semantic profile is compared to three prototypical semantic profiles, representing the three semantic classes of Action, Object, and Substance. The prototypical semantic profiles are built by averaging over the semantic profiles of a sample set of 30 verbs (for Action), countable nouns (for Object), and uncountable nouns (for Substance). We select as our sample words only those that do not appear in our training or test data. The most similar prototypical semantic profile is taken as the model's prediction of the semantic class of a novel word. We then label a semantic prediction by the model as correct if it predicts the class Action for a frame with a verb as its Head, the class Substance for a frame with a noun as its Head whose left Cooc feature is *some*, or class Object for a frame with a noun as its Head whose left Cooc feature is *a/an*. We perform semantic prediction for 150 test frames containing novel words, 50 frames from each category of verbs, countable nouns, and uncountable nouns. Table 2 reports the number of correct answers by our model for each group of test words separately.

Each row of Table 2 shows the number of frames of a specific syntactic type assigned by our model to different semantic categories. According to Brown's experiment children easily distinguish verbs from nouns. Similarly in our model, we observe that all verbs are correctly recognized, i.e., in the last row of the table all verbs are shown to be from the semantic class Action. Prediction for countable nouns stands in the second position of accuracy; with only few numbers misclassified as either Substance or Action. The worst result is observed for Uncountable nouns that were mostly assigned to the Object semantic category. This might be the result of either the appearance of *some* before both countable and uncountable nouns in our training data, or a weakness in the semantic features representing mass entities. We have investigated the utterances in the corpus and found many cases in which the caregiver or mother spoke ungrammatically, e.g., using the determiner *a* before an uncountable noun or using *some* to introduce a countable noun, which both might be partially responsible for the misclassifications. Nonetheless, this result resembles that found by Samuelson and Smith [26]: the authors in this study examine a corpus of CDS and also perform a test on children to study the relation between the syntactic categories and the ontology of words. They also find that predictions from semantic features of entities (such as solidity or shape of objects) to syntax are stronger overall than are predictions from syntax to such semantic features. Moreover, they suggest that children may know that solid objects are named by properties such as shape but not know anything systematic about how categories of non-solid entities are organized.

8 Conclusions and Future Directions

We have used a modified version of an existing categorization algorithm (that of Alishahi and Chrupała [1]) to study the acquisition of syntactic categories in children, and to examine the effect of different types of cues on this process. We have introduced a novel word categorization task, which is an appropriate framework to evaluate the usefulness of language-internal (e.g., co-occurrence) as well as language-external features (e.g., semantic properties), independently from the identity of the word being categorized (i.e., the head word). For example, our results indicate that the categorization of closed-class words (such as determiners) strongly relies on the head word, whereas open-class words (e.g., verbs and nouns) can be successfully categorized based on a combination of syntactic and morpho-/phonological properties, even without taking into account the word itself. In a more detailed investigation of the different cues, we observe that verbs are better recognized when we use phonological or morphological properties in combination with the context (co-occurring words). For nouns, however, using the context alone results in a more precise categorization.

Previous work has shown that, in general, the head word is very effective in categorizing words (see, e.g., Chang et al. [6]). Evaluating the effect of different cues in word categorization models thus needs much care. Studies such as those of Parisien et al. [23] and Alishahi and Chrupała [1] have reported the capability of co-occurrence information in categorizing words. They include, however, the head word itself as part of their features used for categorization. These studies evaluated the performance of their models on various tasks, such as noun/verb disambiguation, and semantic feature prediction. But they did not provide a comparison between their models and a categorization model that only uses the head word. Experiments such as our novel word categorization, or a comparison with a baseline model that only uses head words, are necessary to understand the real effect of different cues in categorization (as also noted by Chrupała and Alishahi [7]). Otherwise, the word identity might obscure the effect of other word properties (see Asr et al. [2] for related experiments).

We have investigated the interaction between semantic properties and the other (morpho-/phonological and syntactic) properties in the formation and use of the word categories. This is done through two semantic prediction experiments, in which the model predicts the semantics of a novel word by drawing on syntactic and morpho-/phonological information about the word. The semantic prediction experiments are meant to investigate how knowledge of morphology/phonology and/or syntax can help the model perform semantic inference. In one experiment, instead of examining whether the assigned category matches the part of speech (PoS) of a novel word, we measure the extent to which the assigned category matches the word based on semantic similarities. A similar experiment has been performed by Chrupała and Alishahi [7], but only considering co-occurrence properties. Results indicate a clear correlation between syntactic (co-occurring words) and semantic properties, but no significant effect when we combine syntactic properties with morphological and/or phonological information. Nonetheless, we should note that some of these findings might be due to the specific semantic features that we have used in our study. Future work will need to further investigate such issues.

In a second experiment, we simulate the experiment of Brown [4], in which young children are asked to predict the semantic class of a novel word, based on its syntactic context—e.g., to predict whether *dax* in "Here is a *dax*" or "Here is *daxing*" refers to an Action or an Object. Results of our simulations indicate that the model—after being trained using syntactic and semantic features of words—can easily recognize a novel verb following the infinitive *to* as referring to an Action. Analogously, a word following the determiner *a/an* is recognized as an Object. These results match those reported by Brown. However, our model does not show similar performance to children when it comes to the recognition of Substances—that is, our model often misclassifies a novel word following *some* as an Object. This might be due to the appearance of both determiners *a/an* and *some* preceding both countable and uncountable nouns in our training data, or due to some weakness

in the semantic features representing mass entities. Nonetheless, these findings are in line with those observed by Samuelson and Smith [26], who find that predictions from semantic features of entities (such as solidity or shape of objects) to syntax are stronger overall than are predictions from syntax to such semantic features. Specifically, Samuelson and Smith suggest that children may know that solid objects are named by properties such as shape but not know anything systematic about how categories of non-solid entities are organized. Future work will need to further look into this issue.

It should be mentioned here that a computational model might exhibit outputs very different from our expectations prior to implementation. For example, our observations indicate that considering the last phoneme of words is not effective for categorization, and this finding is in contrast to that of Onnis and Christiansen [22]. Although ending phonemes are expected to be good approximations of inflectional word suffixes (notably for verbs and adjectives as verified in our experiments), we argue that two issues in our modeling framework possibly inhibit their helpful role: incrementality of the categorization algorithm, and the simultaneous use of several types of cues. During training, a new word such as *string* might be clustered either in a category of nouns based on co-occurrence features (e.g., a preceding determiner *a*), or in a category of verbs based on its ending phoneme (*-ing*). If this word token is mistakenly assigned to a category with some other verbs, this category might in future attract other nouns because of their distributional or semantic similarity to this noisy cluster; hence, the homogeneity of the categories may be threatened. Future research will need to look into better mechanisms for handling noise by an incremental categorization algorithm.

The computational setup presented in this paper can be extended in a number of ways, such as including other types of information that are known to play a role in word categorization by children. The morpho-/phonological set of features that we used here includes the boundary phonemes as well as the word length (number of syllables and letters). Monaghan et al. [20] presented analyses of several other phonological features available in child-directed speech (CDS) that can be taken into account. On the other hand, selection of semantic information still remains controversial and needs much more investigations. We made use of a pre-processed lexicon by Fazly et al. [9] to attribute words with semantic features. However, a more accurate modeling of category learning would extract semantic information from the target CDS corpus (information in the natural learning environment of children). We can also extend our evaluation, e.g., by providing simulations of other relevant psyholinguistic experiments on children (e.g., those of Berko [3] and Samuelson and Smith [26]). An alternate evaluation method has been proposed by Chang et al. [6], which makes use of child utterances in a corpus. Their suggested method is to compare what a computational model has learned to what children learn. This evaluation strategy is particularly applicable in comparative studies on individuals, and can be incorporated into studies such as the one presented here (Tables 3 and 4).

Appendix

Table 3 %Accuracy of novel word categorization in five conditions, similarity threshold set for <300 clusters)

	#of Clusters	Overall	Open-class	Closed-class	Noun	Verb	Adj	Det
Cooc	288	67	70	9	80	66	10	4
Cooc + Morph	258	64	66	10	71	71	19	9
Cooc + Phon	264	73	75	12	73	81	8	18
Cooc + Sem	247	88	90	43	96	93	57	77
Cooc + Morph + Phon + Sem	254	75	77	38	86	79	30	70

Table 4 %Accuracy of novel word categorization in five conditions, similarity threshold set for <500 clusters)

	#of Clusters	Overall	Open-class	Closed-class	Noun	Verb	Adj	Det
Cooc	500	71	73	18	87	73	10	9
Cooc + Morph	497	76	70	8	77	73	22	9
Cooc + Phon	483	74	77	16	84	82	10	13
Cooc + Sem	346	83	85	49	98	82	61	79
Cooc + Morph + Phon + Sem	457	78	80	37	91	80	42	65

References

1. Alishahi, A., & Chrupała, G. (2009) Lexical category acquisition as an incremental process. In *CogSci-2009 Workshop on Psychocomputational Models of Human Language Acquisition*, Amsterdam.
2. Asr, F. T., Fazly, A. & Azimifar, Z. (2010). The effect of word-internal properties on syntactic categorization: A computational modeling approach. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland, USA.
3. Berko, G. J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.
4. Brown, R. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology, 55*(1), 1–5.
5. Cartwright, T., & Brent, M. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*(2), 121–170.
6. Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research, 9*(3), 198–213.
7. Chrupała, G., & Alishahi, A. (2010). Online entropy-based model of lexical category acquisition. In *Proceedings of 14th Conference on Computational Natural Language Learning (CoNLL)* (pp. 182–191), Uppsala, Sweden.
8. Clark, A. (2000). Inducing syntactic categories by context distribution clustering. In *Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference*

on Computational Natural Language Learning (Vol. 7, pp. 91–94). Morristown: Association for Computational Linguistics.
9. Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, Washington, DC.
10. Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: The MIT press. ISBN 026206197X.
11. Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development, 55*(4), 1535–1540.
12. Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*(02), 249–268.
13. Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21–54.
14. Harm, M. (2002). *Building large scale distributed semantic feature sets with WordNet* (Tech. Rep. No. PDP. CNS. 02.01). Carnegie Mellon University, Center for the Neural Basis of Cognition, Pittsburgh, PA.
15. Kaplan, F., Oudeyer, P., & Bergen, B. (2008). Computational models in the debate over language learnability. *Infant and child development, 17*(1), 55–80.
16. Kemp, N., Lieven, E., Tomasello, M. (2005). Young children's knowledge of the "determiner" and "adjective" categories. *Journal of Speech, Language, and Hearing Research, 48*(3), 592–602.
17. Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania, Philadelphia.
18. MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*, volume 2: The database (3rd ed.). MahWah: Lawrence Erlbaum Associates.
19. Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91–117.
20. Monaghan, P., Christiansen, M., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*(4), 259–305.
21. Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*, 357–374.
22. Onnis, L. & Christiansen, M. (2008). Lexical categories at the edge of the word. *Cognitive Science, 32*(1), 184–221.
23. Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning* (pp. 89–96). New York: Association for Computational Linguistics.
24. Pearl, L. (2009). Using computational modeling in language acquisition research. *Experimental Methods in Language Acquisition Research*, 163–184.
25. Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425–469.
26. Samuelson, L. & Smith, L. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition, 73*(1), 1–33.
27. Schütze, H. (1995). Distributional part-of-speech tagging. In *Proceedings of the Seventh Conference on European Chapter of the Association for Computational Linguistics* (pp. 141–148). San Francisco: Morgan Kaufmann Publishers Inc.
28. Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language, 28*, 127–152.
29. Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, 20*(1), 6–10.

In Learning Nouns and Adjectives Remembering Matters: A Cortical Model

Alessio Plebe, Vivian M. De la Cruz, and Marco Mazzone

Abstract The approach used and discussed here is one that simulates early lexical acquisition from a neural point of view. We use a hierarchy of artificial cortical maps that builds and develops models of artificial learners that are subsequently trained to recognize objects, their names, and then the adjectives pertaining to their color. Results of the model can explain what has emerged in a series of developmental research studies in early language acquisition, and can account for the different developmental patterns followed by children in acquiring nouns and adjectives, by perceptually driven associational learning processes at the synaptic level.

1 Introduction

The endeavor of mapping words to meaning is one of the many fundamental linguistic tasks infants are confronted with on their way to becoming full-fledged members of their linguistic community. The challenges learning new words present to the growing child are in no way trivial and neither are those it poses to researchers who are interested in understanding how it happens. Infants, like the linguist in the hypothetical land of the "gavagai" [64], have to figure out how the sounds they hear are related not only to one another, but to the living things, objects and actions in the world around them.

A. Plebe (✉) · V.M. De la Cruz
Department of Cognitive Science, University of Messina, Messina, Italy
e-mail: alessio.plebe@unime.it; vdelacruz@unime.it

M. Mazzone
Laboratory of Cognitive Science, University of Catania, Catania, Italy
e-mail: mazzonem@unict.it

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_5,

In the understanding of this issue, simulation by way of artificial neural networks may prove very useful. We have already introduced in [62] a model based on a hierarchy of maps simulating cortical processes of self-organization, in order to analyze the emergence of object names by visual and acoustic perceptions only. This model has proven apt in accounting for characteristics observed in child development, such as the phenomenon which has been referred to in the literature as "fast mapping", without any need for specialized word-learning mechanisms. In the present work, we extend our model by also introducing color names in combination with the previously used object names. This makes it possible for us to address two related issues. First, learning adjectives has proven to be a more difficult task than learning nouns. Second, syntactic features seem to play a role in learning color terms and other adjectives. We aim to show that both these aspects can be accounted for within our model. The model initially learns object labels (nouns) with accuracy but not words related to a property of the object (such as color adjectives). Then, the model demonstrates improved accuracy in learning words pertaining to properties of objects (adjectives of color) once working memory is potentiated – something that really happens in child and language development as the brain matures. However, this improved accuracy appears to be sensitive to word order, thus showing the emergence of an embryonic syntax. In practice, we will proceed in the following way. In Sects. 1.1 and 1.2, we will survey the literature on learning nouns and adjectives respectively, while in Sect. 1.3 we will consider the other computational models attempting to simulate the first emergence of language. Section 2 will be devoted to the description of our model. In Sect. 3 we will focus on how our model may account for nouns and adjectives acquisition.

1.1 On Learning First Words

How do children go about learning their first words? There is much theoretical debate and contradictory empirical results. Different explanations have been proposed to explain the processes behind the learning of first words with some researchers suggesting that it should be considered more a process than an event [75]. In fact, the way infants learn to recognize objects and the words associated to them undergo radical change across development, with initial stages utilizing mechanisms that differ from those used in later stages of the word learning process [34, 57].

Pinpointing, what all the precise mechanisms involved are, however, continues to be a subject of intense debate and explanations cover a wide spectrum (see [12, 32], for overviews). An in depth review of the competing theories is beyond the scope of this work, but in an attempt to provide a glimpse of the compelling issues involved in the study of language acquisition, which inevitably serve as the context in which our model is placed, we will briefly cover some of the more pertinent ones.

In addition to the researchers, mentioned above, there are those that see early language and conceptual development as being constrained by a series of innate representational biases that assist infants in understanding what to focus on first in their interactions with the objects in the world and the labels that adult speakers use to refer to them. The whole object bias [35,48] and the shape bias [45] are examples of these. Researchers have also proposed that there are innate interpretational constraints that come into play such as mutual exclusivity or an expectation that every object has only one label that will apply to it [49] and the principle of contrast [19], which states that any difference in the form of a word of a language marks a difference in meaning.

There are others that claim that infants initiate the process of language learning equipped with general but powerful expectations of finding commonalities (e.g. category based, property based, or action based) in the things they see, hear or experience, that help them link, linguistic, perceptual and conceptual units [13]. Others yet, have proposed theories in which a child is seen as bringing to the language learning process a whole series of conceptual and cognitive capacities that include different types of core knowledge about animate and inanimate objects and the principles that govern them [14, 18] and as using joint attention mechanisms [77] as well as pragmatic and social cues such as eye gaze and intentionality [30]. Researchers, however, have demonstrated that while 10-month-old infants are sensitive to social cues, they cannot recruit them for word learning and therefore, suggest that at this age infants presumably have to learn words on a simple associative basis [63]. It is not by chance, it seems, that early vocabulary is made up of the objects infants most frequently see [28].

Those with a strictly associative approach see the learning of first words as not being different to other types of learning and as being strongly influenced by learning from perceptual and sensory experience [65, 74]. Early word-learning, according to a number of these researchers can be explained, at least at a certain point in development, by associational learning strategies alone [73]. The processes invoked are not language specific but of a domain-general nature and emerge as a result of development. Others yet, focus on the statistical and computational resources infants might be bringing to the process of word learning [23, 69]. Recent proposals, such as the one found in the current volume [4], simulate with a computational model, behavioral study results that suggest that young children show an early sensitivity to the syntactic and morpho/phonological properties of words and make use of the combination of cues found in their language environment to build their syntactic knowledge of linguistic categories.

The process of learning labels of objects and adjectives referring to their properties (and in particular, color) that we explore in our model, is consistent with an associative approach, but one that explores associative learning at the neural level, something that we will describe in more detail below. Before that, however, we will briefly describe several interesting characteristics of noun and adjective learning that have emerged from behavioral studies with children.

1.2 On Learning Nouns

Word learning in children starts off slow but quickly takes off. Fast-mapping has been claimed to be one of the processes behind the way very young children are able to quickly learn words on the basis of very few exposures. On the average, children start producing their first words by the age of 10–14 months. Shortly after, by the age of two, they are producing approximately 300. At the age of six, their receptive vocabularies can contain over 14,000 words [17].

Cross-linguistic studies have shown that the vocabularies of very young children from a variety of linguistic backgrounds are made up to a large extent by count nouns [7]. Interesting results have emerged from a more recent cross-linguistic study [15] that investigated both the composition of the vocabularies of 20 month-olds, as well as how word classes in their vocabularies co-vary. The children that participated in the study came from seven different contrasting lingustic communities (Argentina, Belgium, France, Israel, Italy, the Republic of Korea, and the United States). This study found that across these languages, children with vocabularies of 51–100 and 101–200 words had more nouns in their productive vocabulary than any other word class (i.e. verbs, adjectives, and closed-class words). Furthermore, those that had 101–200 word vocabularies, said more verbs than adjectives or closed-class words, and more adjectives than close-classed words. In larger vocabularies, (201–500 words), the same pattern was found, with nouns being by far the most spoken words, but with all the other word classes being positively correlated. Said differently, the pattern found in children with large vocabularies was that nouns, verbs and adjectives develop in tandem, and not in competition with one another. Children with very small-vocabularies (0–50 words), that had just begun word-learning, on the other hand, said more nouns than adjectives or closed-class words, but no differences emerged between nouns and verbs. This pattern of co-variation does not seem to be specific to culture in these languages, in that it continued to emerge under different conditions. In sum, the major findings that emerged from this study is that there is a bias for nouns in the early vocablularies of children learning the languages investigated, and that except for the case of very small vocabularies, the differentiation among the classes of words increases with vocabulary size.

A variety of explanations have been proposed to explain what seems to be a universal noun advantage and the apparent gap between the learning of nouns as opposed to other gramatical forms such as adjectives. Some of these accounts have postulated highly specific innate mechanisms or constraints such as the ones discussed in the previous section.

The cross-linguistic study described above [15] for example, interprets the noun bias that emerged in the languages they investigated as being indicative of the important role constraints play in child lexical development. These constraints however, would necessarily have to be flexible enough to be modifiable by the morphology, saliency, frequency and pragmatics of words in different languages, something they see as being in line with the emergentist coalition model of the

origns of word learning [36], that views word learning as an emergent product of multiple factors that would include cognitive constraints, social-pragmatic factors and global attential mechanisms.

Others have claimed a purely logical explanation to the noun advantage over other grammatical classes. Gentner [27] for example, proposed that nouns are easily learned because they refer to entities, and not relations between entities, like adjectives and verbs. Yet another account, focuses on the fact that nouns are more readily learned because other grammatical classes such as verbs and adjectives are predicates and depend on nouns for their meaning. Their learning is necessarily grounded on the acquisition of nouns first [54].

1.3 On Learning Adjectives

Children's acquisition of adjectives on the other hand, is slow and their use is prone to errors. The learning of color terms seem to be particularly challenging early on in development, with very young children seeming to be almost incapable of learning them. Charles Darwin, himself, noting the lack of color terms used by his own child, mistakingly speculated that children are initially born color blind [21].

Some literature on this phenomenon has been reviewed by Gasser and Smith [26], which points to three kinds of evidence. First, nouns dominate early productive vocabularies of children, while adjectives are rare or nonexistent: for instance, in Nelson's study of 18 children learning English, fewer than 7 % of the first 50 words were adjectives. Second, experimental studies of word learning show that the application of a novel adjective appears more slowly and more variably determined than the application of names for things. Third, there is some evidence that children are more prone to errors with adjectival than with nominal meanings.

As has already been mentioned in the above section on noun learning, another reason why adjectives (and verbs) may also be harder to learn may be due to purely logical reasons, such as the fact that they refer to entities and not to the relations that exist between them. We believe that there is no reason to invoke special mechanisms in explaining the temporal gap between adjective and noun learning and that they can be explained, at least in an initial early phase, by way of the associations between visual and acoustic stimuli experienced by young children. Our model supports this thesis. We think that in the process of acquiring adjectives two factors are initially involved.

The first is the poor covariation of features. The unidimensional properties referred to by adjectives do not covary systematically with other features. Psychologists have observed that young children have difficulty attending selectively to individual dimensions. This problem does not affect noun learning, since common nouns label objects that are similar across many inter-related and correlated properties. This fact has a clear neuro-computational grounding: artificial and natural neural networks are based on the simple detection of co-occurrences, and therefore, the advantage of nouns over adjectives admits a straightforward

associative explanation. The consideration of covariation not only may explain the difficulty in learning adjectives; it also suggests that giving a name to unidimensional visual properties may enhance the ability to selectively attend to them, thanks to the covariation between acoustic and visual properties.

This leads us to the second factor: adjectives are normally heard by children together with other linguistic items and are one of the first syntactic challenges novice language learners face. Utterances with the sequence `[Adj Noun]` are what we like to refer to as, "embryonic syntax", and it departs from the single sound pattern to reference scheme initially experienced by infants. The explanation based on these two points, fits well with developmental evidence [70] that the learning of adjectives, while difficult at first, then gets easier, once children have acquired more knowledge about their language.

The ability to acquire adjectives very likely also depends on the maturation of brain circuits, especially in the prefrontal cortex, a sustained representational device [1, 25]. In fact, language development crucially depends on the development of an expanding working memory capacity, the kind of short-term memory theorized by Baddeley [6], which would pave the way to the processing of complex sequences of sounds in that it consents the retention of the meanings of sounds as they are encountered. The emergence of syntactic processes, such as being sensitive to the order in which words appear, would also depend on these potentiated memory circuits found in the temporoparietal and prefrontal areas, known to develop slowly in ontogeny. This would account for why more complex grammatical forms are acquired later in development: they depend on an expanded memory capacity that is just not available in early infancy. The work done with our model has reflected this. Less memory is necessary for learning nouns initially, but adjective learning is made possible and subsequently easier, only once memory circuits have been potentiated.

1.4 Modeling Noun and Adjective Acquisition

Several attempts have been made in modeling computationally different aspects of children's language learning. Models consent, the zeroing in on particular aspects of processes that are difficult to separate in experiments with real children, but their limits have been related to the biological plausability of the mathematical approaches, and or the realism of the stimuli.

Only one model [67] seems to be able to deal with real stimuli, in learning words of visual objects. The system used recordings of utterances of caregivers as they spontaneously interacted with their infants while playing with objects, and real images of these objects. The model was able to segment words from utterances and to associate the proper word with the object seen with impressive accuracy, demonstrating that early word learning can be based on co-occurrence patterns within the visual context. The computations implemented in this model, however, are careful combinations of standard image processing and signal processing algorithms, without any relationship to biological brain computations.

Inside the classical PDP framework [68], an abstract neural model [66] explored the emergence of simple conceptual systems in infancy. Their model learned categories of birds, fish, flowers and trees, by associating a predefined set of visual features, like "red" or "branches", with a fixed set of attributes, such as "can walk", "is living" and so on. This model is interesting in that it demonstrates that semantic categories of objects and attributes can emerge without appealing to cognitive constraints, as an effect of mere statistical regularities. However, the unrealistic modeling of both stimuli and cognitive architecture limits its explanatory value.

Recently, similar approaches have explored specific aspects, such as fast-mapping [51]. This model is based on standard self-organizing maps [43], and reproduces interesting aspects of language learning, like slips of the tongue and mispronunciation effects, but lacks correspondence between its mathematics and how brain computation is structured.

Our adjective and noun learning model is an attempt to build a system that adheres in varying degrees to the reality of the corresponding computations taking place in the brain, dealing with realistic inputs. As far as we know, this is the first model combining visual and auditory paths by simulated cortical maps. It is based on an architecture developed by the authors, that began with object recognition [61], progressed in combining the visual and auditory pathways [62], and here, is further extended to reproduce the emergence of embryonic syntax.

Apart from the details of the way we structure the visual and auditory pathways (see Sect. 2), a general feature which makes our model biologically realistic to a significant degree is its hierarchical organization of cortical maps. This hierarchical architecture is of the sort recommended by Mayor and Plunkett [51] as a potential solution to difficulties encountered within their model, in particular the "inability to learn new words after the visual and auditory maps have stabilized" [51, p. 20]. In their opinion, this could be overcome by employing "hierarchies (or heterarchies) of maps in both the visual and auditory pathways of the model, (so as to mimic) aspects of the organization of visual and auditory cortex". This is precisely one virtue of our model.

2 Description of the Model

This section will describe the model in detail and the rationale behind the choices of its design. One of the major decisions in designing neural models is the trade-off between the biological details taken into account in the simulation, and the complexity of the overall system. Our goals concern the simulation of language, which is a top level cognitive function, involving most of the brain. More specifically, we are interested in the early development of word recognition based on perceptually salient objects or properties, a phenomenon strongly related to synaptic changes at the cortical level.

2.1 Basic Units of the Model

According to the design principle stated above, one basic structure of the model is a sheet of artificial neurons, that correspond to the neurophysiological concept of "cortical maps", originally used by Mountcastle [55], and widely used in vision science as a well grounded way of partitioning the cortex [82, 83]. The extent to which the correspondence between modules in our model and cortical maps is faithful, is a matter of degree. There is a large gap in knowledge between what is known about the visual system, and what is instead known of the auditory stream. The boundaries where the two systems meet are even more obscure.

Inside each cortical sheet unit, the most basic unit is the artificial neuron, that behaves according to the LISSOM scheme (*Laterally Interconnected Synergetically Self-Organizing Map*) [72]. This artificial neuron is not the equivalent of a single cell, and details of the electrical activity of the neuron in real time are not simulated. It is instead the equivalent of a vertical columnar assembly of cells, whose firing rate is assumed by an average activation level. The simple fact that one LISSOM neuron could therefore represent thousands of biological neurons, is a clear sign that models are still far from being a precise reproduction of real brain computations, our understanding of which is still limited, especially for higher cognition. Though not being strictly equivalent to a single neural cell, the LISSOM neuron possesses two important aspects of cortical circuits: the modifiable lateral connections of excitatory and inhibitory types, and Hebbian-based plasticity.

The basic equation of the LISSOM describes the activation level x_i of a neuron i at a certain time step k:

$$x_i^{(k)} = f\left(\frac{\gamma_A}{1+\gamma_N \mathbf{I}\cdot\mathbf{v}_{r_A,i}}\mathbf{a}_{r_A,i}\cdot\mathbf{v}_{r_A,i} + \gamma_E \mathbf{e}_{r_E,i}\cdot\mathbf{x}_{r_E,i}^{(k-1)} - \gamma_H \mathbf{h}_{r_H,i}\cdot\mathbf{x}_{r_H,i}^{(k-1)}\right), \quad (1)$$

where the vectors $\mathbf{x}_{r_E,i}^{(k-1)}$ and $\mathbf{x}_{r_H,i}^{(k-1)}$ are the activations of all neurons in the map, where a lateral connection exists with neuron i of an excitatory or inhibitory type, respectively. Their fields are circular areas of radius, respectively, r_E, r_H. Vectors $\mathbf{e}_i$ and $\mathbf{h}_i$ are composed by all connection strengths of the excitatory or inhibitory neurons projecting to i. The vectors $\mathbf{v}$ and $\mathbf{x}_i$, as before, are the input and the neural code. The scalars γ_X, γ_E, and γ_H, are constants modulating the contribution of afferents.

The scalar γ_N controls the setting of a push-pull effect in the afferent weights, allowing inhibitory effects without negative weight values. Mathematically, it represents dividing the response from the excitatory weights by the response from a uniform disc of inhibitory weights over the receptive field of neuron i. Vector $\mathbf{I}$ is just a vector of 1's of the same dimension of $\mathbf{x}_i$. The function f can be any monotonic non-linear function limited between 0 and 1. For computational economy, it has been implemented as a piecewise linear approximation of the sigmoid function.

The final activation value of the neurons is assessed after a certain settling time K, typically about ten time steps. All connection strengths adapt according to the general Hebbian principle, and include a normalization mechanism that counterbalances the overall increase of connections of the pure Hebbian rule. The equations are the following:

$$\Delta \mathbf{a}_{r_{\mathrm{A}},i} = \frac{\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}}{\|\mathbf{a}_{r_{\mathrm{A}},i} + \eta_{\mathrm{A}} x_i \mathbf{v}_{r_{\mathrm{A}},i}\|} - \mathbf{a}_{r_{\mathrm{A}},i}\,, \tag{2}$$

$$\Delta \mathbf{e}_{r_{\mathrm{E}},i} = \frac{\mathbf{e}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}}{\|\mathbf{a}_{r_{\mathrm{E}},i} + \eta_{\mathrm{E}} x_i \mathbf{x}_{r_{\mathrm{E}},i}\|} - \mathbf{e}_{r_{\mathrm{E}},i}\,, \tag{3}$$

$$\Delta \mathbf{h}_{r_{\mathrm{H}},i} = \frac{\mathbf{h}_{r_{\mathrm{H}},i} + \eta_{\mathrm{A}} x_i \mathbf{x}_{r_{\mathrm{H}},i}}{\|\mathbf{h}_{r_{\mathrm{H}},i} + \eta_{\mathrm{A}} x_i \mathbf{x}_{r_{\mathrm{H}},i}\|} - \mathbf{h}_{r_{\mathrm{H}},i}\,, \tag{4}$$

where $\eta_{\{\mathrm{A,E,H}\}}$ are the learning rates for the afferent, excitatory, and inhibitory weights.

The overall model is a combination of artificial cortical sheets that reproduce essentially the part of the brain involved in learning nouns and adjectives of perceptual salience: the visual and the auditory pathways. Each of the two processing streams is fed by realistic stimuli, and therefore also comprises, in an approximate simulation, the components that transduce the external signals, and the subcortical components.

2.2 The Visual Pathway

An outline of the modules that make up the visual pathway is shown in Fig. 1a, it is built upon and extends a previous model of visual object recognition [61]. The visual system encompasses the ventral stream only, the "what" stream in the classical dichotomy established by Ungerleider and Mishkin [78], specialized for object recognition [59]. The governing equations in this section of the model are the following:

$$x^{(\odot)} = f\left((\mathbf{l}_{r_{\mathrm{A}}} + \mathbf{m}_{r_{\mathrm{A}}}) \cdot \left(\mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})} - \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})}\right)\right) \tag{5}$$

$$x^{(\circledcirc)} = f\left((\mathbf{l}_{r_{\mathrm{A}}} + \mathbf{m}_{r_{\mathrm{A}}}) \cdot \left(\mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})} - \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})}\right)\right) \tag{6}$$

$$x^{(\mathrm{R}^{+}\mathrm{G}^{-}\odot)} = f\left(\mathbf{l}_{r_{\mathrm{A}}} \cdot \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})} - \mathbf{m}_{r_{\mathrm{A}}} \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})}\right) \tag{7}$$

$$x^{(\mathrm{R}^{+}\mathrm{G}^{-}\circledcirc)} = f\left(\mathbf{l}_{r_{\mathrm{A}}} \cdot \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})} - \mathbf{m}_{r_{\mathrm{A}}} \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})}\right) \tag{8}$$

$$x^{(\mathrm{G}^{+}\mathrm{R}^{-}\odot)} = f\left(\mathbf{m}_{r_{\mathrm{A}}} \cdot \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})} - \mathbf{l}_{r_{\mathrm{A}}} \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})}\right) \tag{9}$$

$$x^{(\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc)} = f\left(\mathbf{m}_{r_{\mathrm{A}}} \cdot \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{W}})} - \mathbf{l}_{r_{\mathrm{A}}} \mathbf{g}_{r_{\mathrm{A}}}^{(\sigma_{\mathrm{N}})}\right) \tag{10}$$

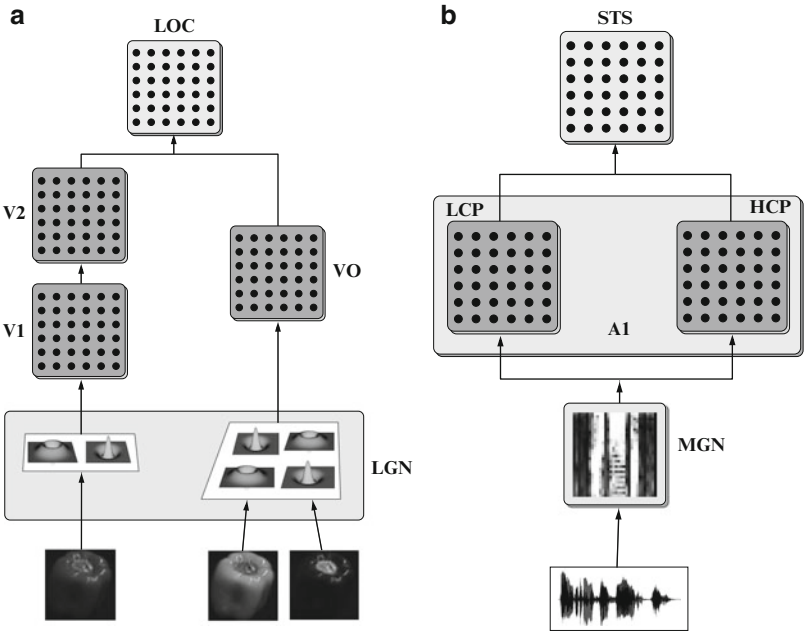

Fig. 1 Scheme of the visual pathway (**a**) and the auditory pathway (**b**) in the model. The visual pathway is composed by LGN (Lateral Geniculated Nucleus), V1 (Primary Visual Cortex), V2 (Secondary Visual Cortex), VO (Ventral Occipital Cortex), LOC (Lateral Occipital Complex). The auditory pathway is composed by MGN (Medial Geniculated Nucleus), A1-LPC (Auditory Primary Cortex – Low-Probability Connections), A1-HPC (Auditory Primary Cortex – High-Probability Connections), STS (Superior Temporal Sulcus).

$$x^{(\mathrm{V1})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{V1})}\left(\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{V1}\leftarrow\odot)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\odot)} + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{V1}\leftarrow\circledcirc)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\circledcirc)}\right)\right.$$
$$\left. + \gamma_{\mathrm{E}}^{(\mathrm{V1})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{V1})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{V1})} - \gamma_{\mathrm{H}}^{(\mathrm{V1})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{V1})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{V1})}\right) \tag{11}$$

$$x^{(\mathrm{V2})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{V2})}\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{V2}\leftarrow V1)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(V1)} + \gamma_{\mathrm{E}}^{(\mathrm{V2})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{V2})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{V2})} - \gamma_{\mathrm{H}}^{(\mathrm{V2})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{V2})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{V2})}\right) \tag{12}$$

$$x^{(\mathrm{VO})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{VO})}\left(\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{VO}\leftarrow\mathrm{R}^{+}\mathrm{G}^{-}\odot)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\mathrm{R}^{+}\mathrm{G}^{-}\odot)} + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{VO}\leftarrow\mathrm{R}^{+}\mathrm{G}^{-}\circledcirc)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\mathrm{R}^{+}\mathrm{G}^{-}\circledcirc)}\right.\right.$$
$$\left.\left. + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{VO}\leftarrow\mathrm{G}^{+}\mathrm{R}^{-}\odot)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\mathrm{G}^{+}\mathrm{R}^{-}\odot)} + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{VO}\leftarrow\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(\mathrm{G}^{+}\mathrm{R}^{-}\circledcirc)}\right)\right.$$
$$\left. + \gamma_{\mathrm{E}}^{(\mathrm{VO})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{VO})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{VO})} - \gamma_{\mathrm{H}}^{(\mathrm{VO})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{VO})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{VO})}\right) \tag{13}$$

$$x^{(\mathrm{LOC})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{LOC})}\left(\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{LOC}\leftarrow V2)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(V2)} + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{LOC}\leftarrow VO)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(VO)}\right)\right.$$
$$\left. + \gamma_{\mathrm{E}}^{(\mathrm{LOC})}\mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{LOC})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{LOC})} - \gamma_{\mathrm{H}}^{(\mathrm{LOC})}\mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{LOC})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{LOC})}\right) \tag{14}$$

There are two distinct pathways, one achromatic, processed by Eqs. (5), (6), and another sensitive to colors, limited here to medium and long wavelengths. The equations are: (7)–(10). The symbol $\odot$ refers to on-center type receptive fields, and symbol $\circledcirc$ to off-center receptive fields. The profile of all visual receptive fields is given by differences of two Gaussian $\mathbf{g}^{(\sigma_N)}$ and $\mathbf{g}^{(\sigma_W)}$, with $\sigma_N < \sigma_W$. This is an approximation of the combined contribution of gangliar cells and LGN [22].

In all equations the activation x has to be taken as the activation of a generic i-th neuron of that level, and all receptive fields have to be intended as referring to that neuron, the index i, and the indication of the radius r of the circular receptive field, has been omitted for clarity. In exploiting the modularity of the model, enacted by the correspondence with cortical maps, a simplification has been introduced by way of the separation of the processing of shape and color. Shape is elaborated through V1 and V2, by Eqs. (11) and (12), and the processing of color is entrusted to VO, with Eq. (13).

There is evidence, in fact, that suggests that in the visual system no segregation of functions such as shape or color processing takes place, and that almost all visual cortical maps cooperate in analyzing form, color, motion and stereo information [71, 79]. On the other hand, it is clear that visual areas are not equally involved in all aspects of object recognition. It is possible to identify specialization in one main function in certain maps. This is the case in what has been called the color center area by Zeki [87, 88], who named it "V4", we are using the more general name of VO (Ventral Occipital), given by Wandell et al. [83]. V1 is the well-known primary visual cortex, the most studied part of the brain [37, 38]. One of its main functions is the organization of the map into domains of orientation tuned neurons [11, 80], which are fundamental for early shape analysis, our model discards the contributions of V1 to all other processes. The main projection from V1 is to its immediately anterior area, V2, whose functions are less understood than V1. It is probably responsible for shape analysis at a level of complexity and scale larger than that of V1 [2, 40, 42]. Equation (14) describes the convergence of shape and color processing paths into the LOC model map, corresponding to the area in the human cortex thought to be crucial for the task of recognition in vision, located anterior to Brodmann's area 19, near the lateral occipital sulcus, extending into the posterior and mid fusiform gyrus and occipital-temporal sulcus, with an overall surface size similar to V1. Perhaps the most important idea, one that has obtained a certain amount of consensus, is that this area is involved in visual behavior in which recognition is the main task [31, 44, 84].

2.3 Auditory Pathway

The auditory path includes the medial geniculate nucleus, the auditory primary cortex, and the superior temporal sulcus. An outline of the modules that make up this pathway is shown in Fig. 1b. The equations are the following:

$$x_{\tau,\omega}^{(\text{◫})} = \left| \sum_{t=t_0}^{t_M} v(t) w(t-\tau) e^{-j\omega t} \right|^2 \tag{15}$$

$$\begin{aligned} x^{(\text{LPC})} = f \Big(& \gamma_{\text{A}}^{(\text{LPC})} \mathbf{a}_{r_{\text{A}}}^{(\text{LPC}\leftarrow\text{◫})} \cdot \mathbf{x}_{r_{\text{A}}}^{(\text{◫})} \\ & + \gamma_{\text{E}}^{(\text{LPC})} \mathbf{e}_{r_{\text{E}}}^{(\text{LPC})} \cdot \mathbf{x}_{r_{\text{E}}}^{(\text{LPC})} - \gamma_{\text{H}}^{(\text{LPC})} \mathbf{h}_{r_{\text{H}}}^{(\text{LPC})} \cdot \mathbf{x}_{r_{\text{H}}}^{(\text{LPC})} \Big) \end{aligned} \tag{16}$$

$$\begin{aligned} x^{(\text{HPC})} = f \Big(& \gamma_{\text{A}}^{(\text{HPC})} \mathbf{a}_{r_{\text{A}}}^{(\text{HPC}\leftarrow\text{◫})} \cdot \mathbf{x}_{r_{\text{A}}}^{(\text{◫})} + \\ & \gamma_{\text{E}}^{(\text{HPC})} \mathbf{e}_{r_{\text{E}}}^{(\text{HPC})} \cdot \mathbf{x}_{r_{\text{E}}}^{(\text{HPC})} - \gamma_{\text{H}}^{(\text{HPC})} \mathbf{h}_{r_{\text{H}}}^{(\text{HPC})} \cdot \mathbf{x}_{r_{\text{H}}}^{(\text{HPC})} \Big) \end{aligned} \tag{17}$$

$$\begin{aligned} x^{(\text{STS})} = f \Big(& \gamma_{\text{A}}^{(\text{STS})} \left(\mathbf{a}_{r_{\text{A}}}^{(\text{STS}\leftarrow LPC)} \cdot \mathbf{x}_{r_{\text{A}}}^{(LPC)} + \mathbf{a}_{r_{\text{A}}}^{(\text{STS}\leftarrow HPC)} \cdot \mathbf{x}_{r_{\text{A}}}^{(HPC)} \right) \\ & + \gamma_{\text{E}}^{(\text{STS})} \mathbf{e}_{r_{\text{E}}}^{(\text{STS})} \cdot \mathbf{x}_{r_{\text{E}}}^{(\text{STS})} - \gamma_{\text{H}}^{(\text{STS})} \mathbf{h}_{r_{\text{H}}}^{(\text{STS})} \cdot \mathbf{x}_{r_{\text{H}}}^{(\text{STS})} \Big) \end{aligned} \tag{18}$$

In (15) the symbol ◫ refers to the spectrotemporal representation of the auditory signal, the horizontal dimension τ is time, and the vertical dimension ω is frequency. Function $w(\cdot)$ in (15) is a short term temporal window, that performs a spectrogram-like response, similar to that given by the combination of cochlear and MGN processes [16].

Very little is known about the computational organization of the auditory primary cortex, compared to the early visual maps [47]. Our model discards binaural interaction, and preserves the main connectivity from single cochlear signals in the Medial Geniculate nucleus to A1. A large body of evidence points to an organization of A1 with a fundamental dependency on sound frequencies along one cortical dimension, and a distribution of neural responses to temporal properties [53, 85]. The auditory primary cortex is simulated by a double sheet of neurons, to take into account a double population of cells found in this area [5], where the so-called LPC (*Low-Probability Connections*) is sensitive to the stationary component of the sound signal and the HPC (*High-Probability Connections*) population responds to transient inputs mainly. Equation (18) states the projection from the primary auditory cortex to STS. This is the model's correlate of a region in the cortical ventral auditory stream, on which there is accumulating evidence and a convergence of opinion on its role in representing and processing phonological information [8, 9, 33, 46].

2.4 The Higher Cortical Map

The model map where the ventral visual path and the auditory path meet is PFC. There are actually several areas where visual and auditory signals converge, and more than one area activated in categorization and syntactic tasks. One reason for

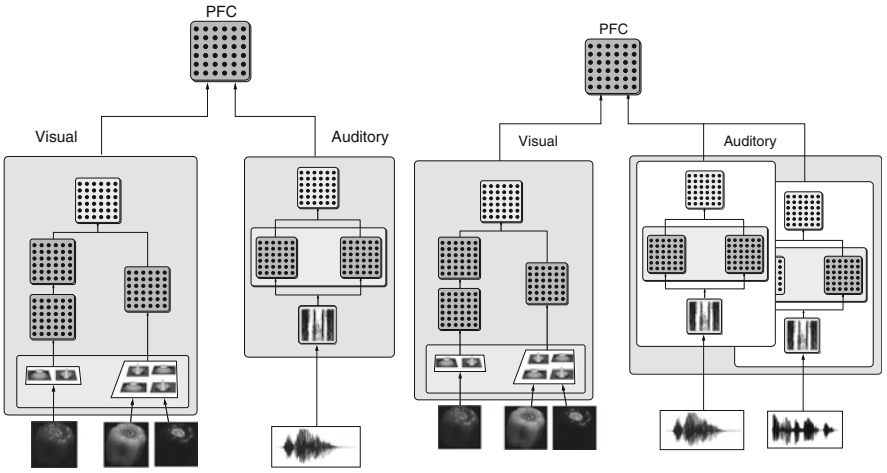

Fig. 2 Scheme of the complete model, in two versions. The one on the *left* has only simple afferent connections to PFC (PreFrontal Cortex). The version on the *right* has additional connections, that enable working memory for syntactic processing

calling the model map PFC, is due to the well established role of the biological lateral prefrontal cortex in object categorization [3, 24, 39, 52, 86]. The main difference with respect to the representation of objects in inferotemporal areas is that lateral PFC could form a more abstract kind of categorization, which is proper to lexical categories.

The model's PFC, in this role of abstract representation of visual and linguistic information, is governed by the following equation:

$$x^{(\mathrm{PFC})} = f\left(\gamma_{\mathrm{A}}^{(\mathrm{PFC})}\left(\mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{PFC}\leftarrow LOC)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(LOC)} + \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{PFC}\leftarrow STS)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(STS)}\right)\right.$$
$$\left. + \gamma_{\mathrm{E}}^{(\mathrm{PFC})} \mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{PFC})} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{PFC})} - \gamma_{\mathrm{H}}^{(\mathrm{PFC})} \mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{PFC})} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{PFC})}\right) \tag{19}$$

The overall scheme of the model, in the case of PFC working as from Eq. (19), is shown in the left of Fig. 2.

There is an additional important function of the prefrontal cortex, that justifies the name given to this model map, and that is essential in the scope of this experiment. Recursive connections between temperoparietal and prefrontal areas are supposed to support the kind of short-term memory theorized by Baddeley [6]. Large cortico-cortical networks are essential in most aspects of language understanding, as in the well-known phonological rehearsal loop. This is of course, only a small part of the role working memory plays. Cortical loops involving the prefrontal cortex [29] allow the ability to keep the semantic meanings of sounds under attention as they

Table 1 Parameters for all neural layers of the model

layer	size	r_A	r_E	r_H	γ_X	γ_E	γ_H	γ_N
LGN	112	2.6	–	–	–	–	–	–
MGN	32	–	–	–	–	–	–	–
V1	96	8.5	1.5	7.0	1.5	1.0	1.0	0
A1	24	3.5	2.5	5.5	5.0	5.0	6.7	0.8
V2	30	7.5	8.5	3.5	50.0	3.2	2.5	0.7
VO	30	24.5	4.0	8.0	1.8	1.0	1.0	0
LOC	16	6.5	1.5	3.5	1.8	1.0	1.5	0
STS	16	3.5	2.5	2.5	2.0	1.6	2.0	0
PFC	24	6.5	4.5	6.5	1.5	3.5	4.1	0

are being formulated and the posing of constraints for the emergence of syntactic processes. In the second version of the model, corresponding to the scheme on the right in Fig. 2, the equation of PFC is the following:

$$
\begin{aligned}
x^{(\mathrm{PFC}^*)} = f\Big(&\gamma_{\mathrm{A}}^{(\mathrm{PFC}^*)} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{PFC}^* \leftarrow LOC)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(LOC)} + \\
&\gamma_{\mathrm{E}}^{(\mathrm{PFC}^*)} \mathbf{e}_{r_{\mathrm{E}}}^{(\mathrm{PFC}^*)} \cdot \mathbf{x}_{r_{\mathrm{E}}}^{(\mathrm{PFC}^*)} - \gamma_{\mathrm{H}}^{(\mathrm{PFC}^*)} \mathbf{h}_{r_{\mathrm{H}}}^{(\mathrm{PFC}^*)} \cdot \mathbf{x}_{r_{\mathrm{H}}}^{(\mathrm{PFC}^*)} \\
&+ \sum_{\zeta=1}^{N_\zeta} \Big(\gamma_{\mathrm{A}}^{(\mathrm{PFC}^*)} \mathbf{a}_{r_{\mathrm{A}}}^{(\mathrm{PFC}^* \leftarrow STS)} \cdot \mathbf{x}_{r_{\mathrm{A}}}^{(STS)_\zeta} \Big) \Big)
\end{aligned}
\tag{20}
$$

where ζ are discrete temporal delays, corresponding to the presentation of spectrograms of progressive words in the sequence of the utterance. In this experiment $N_\zeta = 2$, since the sentence is the sequence `[Adj Noun]`.

We used the simplest model as corresponding to early stages in development, at the onset of language acquisition, around 9–12 months of age, and the model with additional connections providing working memory abilities, at a more mature stage of development, corresponding to about 14–20 months of age.

The Table 1 summarizes the values of the main parameters in the equations here described, for all the maps in the model.

3 Nouns and Adjectives Acquisition

In this section we will describe how the two models have been trained in this experiment, and report on the functions developed in the cortical maps of the models. For lack of space a short account will be given of the functions of the lower cortical maps, referring to other works for further details. We will focus, instead, on the linguistic abilities that emerged in the PFC maps.

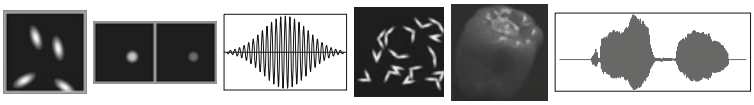

Fig. 3 Example of stimuli to the model. From the *left*, elongated blobs input to V1, hue circular blobs for VO, wave trains for A1, couple of blobs for V2, real images for the visual path, and word waves for the auditory path

3.1 Simulation of Intrinsic and Extrinsic Experience

The model has been exposed to a variety of stimuli, in different stages of its development, that to various extents parallel periods of human development from the pre-natal stage to that of early language acquisition. Initially only V1, VO and A1-LPC, A1-HPC maps are allowed to modify their synaptic weights, by Eqs. (2)–(4). The stimuli presented to V1 and VO are synthetic random blobs that mimic waves of spontaneous retinal activity, that are known to play a fundamental role in the ontogenesis of the visual system [41,50,76]. Blobs presented to V1 are elongated along random directions, to stimulate orientation selectivity. Blobs to VO are circular, with constant hues, and random size, position, and intensity. The A1 maps are exposed to short trains of waves sweeping linearly around a central frequency. Time durations, central frequencies and sweeping intervals are changed randomly. The next period of development involves the V2 and STS maps. The visual stimuli comprises pairs of elongated blobs with a coinciding end point, to enhance the experience of patterns that are slightly more complex than lines, such as corners. The auditory stimuli are synthesized waves of the 7,200 most common English words (from http://www.bckelk.uklinux.net/menu.html), with length of three to ten characters. All words are converted from text to waves using *Festival* software [10], with cepstral order 64 and a unified time window of 2.3 s. In the development stage that corresponds to that just after eye opening, natural images are used. In order to include the primary and most realistic difficulty in recognition, which is the identification of an object under different views, the COIL-100 collection has been used [56], where for each of the 100 objects, 72 different views are available. In most experiments, unless otherwise stated, only eight views per object have been used during the learning phase of the model, and all 72 views are used in the testing phases.

The last stage of the experiment simulates events in which an object is viewed and a word corresponding to its basic category is heard contemporaneously. The 100 objects have been grouped manually into 38 categories. We deliberately used some categories that do not have strictly perceptual traits, such as `medicine`, that make the task of gathering exemplars in the same category particularly difficult, due to the lack of cues regarding the purpose of `medicine`, in order to simulate the challenges infants are faced with when trying to map new words to meanings.

Examples of the stimuli used can be seen in Fig. 3.

3.2 Emergence of Organization in the Lower Maps

At the end of development, different types of organization are found in the lower maps that enable the performance of processes that are essential to recognition, and that are similar to those found in corresponding brain areas. The model's V1 map organized orientation selectivity, with responsiveness of neurons to oriented segments arranged over repeated patterns of gradually changing orientations, broken by few discontinuities, resembling one known to be found in biological primary cortex [11, 80]. In the VO map of the model, most neurons respond to specific hues, regardless of intensity. This is one of the basic features of color processing. Color constancy is crucial in object recognition and is known to develop somewhere between 2 and 4 months of age [20]. The kind of mapping found in A1 is typically tonotopic, and it encodes the dimensions of frequency and time sequences in a sound pattern. This is known to be the main ordering of neurons in biological A1 [81]. The main organization found in the V2 map is responsiveness to angles, especially in the 60° and 150° range. This kind of selectivity is one of the major phenomena recently discovered in biological V2 [2, 40]. In the model's LOC map most neurons exhibit invariant responses to objects. Invariance, the ability to recognize known objects despite large changes in their appearance on the sensory surface, is one of the most important properties to have in an object-recognition cortical area. It has been identified in human LOC by several studies [31, 44, 84].

We refer to other published works for details on the functions that emerged in V1, VO [61] V2 [58], LOC [60] and STS [62].

3.3 Representation of Nouns and Adjectives in Model PFC

It is in the upper PFC map where we expect categorization to take place that concerns both visual and word forms. We have a PFC map in the immature model ruled by Eq. (19), and the starred PFC map in the mature, working memory equipped model, ruled by Eq. (20). For both, a common method of analysis has been carried out, for analyzing the distributions of neural activation as population coding of categories. Let us introduce the following function:

$$x_i(s) : S \in \mathcal{S} \rightarrow \mathbb{R}; \quad s \in S \in \mathcal{S}, \tag{21}$$

that gives the activation x of a generic neuron i in the PFC or PFC* maps, in response to the presentation of the stimulus s to the system. This stimulus is an instance of a class S, belonging to the set of all classes of stimuli $\mathcal{S}$ available in the experiment. For a class $S \in \mathcal{S}$ we can define the two sets:

$$X_{S,i} = \left\{ x_i(s_j) : s_j \in S \right\}; \quad \overline{X}_{S,i} = \left\{ x_i(s_j) : s_j \in S' \in (\mathcal{S}/S) \right\}. \tag{22}$$

Therefore the set $\overline{X}_{S,i}$ includes values of neuron i in responses to all possible stimuli not belonging to S. We can then associate to the class S a set of neurons in the map, by ranking it with the following function:

$$r(S,i) = \frac{\mu_{X_{S,i}} - \mu_{\overline{X_{S,i}}}}{\sqrt{\frac{\sigma_{X_{S,i}}}{|X_{S,i}|} + \frac{\sigma_{\overline{X_{S,i}}}}{|\overline{X}_{S,i}|}}}, \tag{23}$$

where μ is the average and σ the standard deviation of the values in the two sets, and $|\cdot|$ is the cardinality of a set. Now the following relation can be established as the population code of a class S:

$$p(S) : \mathcal{S} \to \{i_1, i_2, \cdots, i_M \ : r(S,i_1) > r(S,i_2) > \cdots > r(S,i_M) > L_r\}, \tag{24}$$

where M, the maximum number of coding neurons, is a given constant, typically one order of magnitude smaller than the number of neurons in the map, and L_r is a threshold for the lowest acceptable ranking for a neuron to be taken as coding for S. The classes of stimuli that can be used in the two models are slightly different.

$$s_N^{(\mathrm{PFC})} = \langle o, n\rangle \in N = \bigcup_{O \in \mathcal{O}_N} O \times \mathcal{U}_N \tag{25}$$

$$s_A^{(\mathrm{PFC})} = \langle o, a\rangle \in A = \bigcup_{O \in \mathcal{O}_A} O \times \mathcal{U}_A \tag{26}$$

$$s_N^{(\mathrm{PFC}^*)} = \langle o, a, n\rangle \in N = \left\{\langle \omega, \alpha, \pi\rangle : \omega \in \mathcal{O}_N \ \wedge \ \alpha \in \mathcal{U}_{A(\omega)} \ \wedge \ \pi \in \mathcal{U}_N\right\} \tag{27}$$

$$s_A^{(\mathrm{PFC}^*)} = \langle o, a, n\rangle \in A = \left\{\langle \omega, \alpha, \pi\rangle : \omega \in \mathcal{O}_A \ \wedge \ \alpha \in \mathcal{U}_\alpha \ \wedge \ \pi \in \mathcal{U}_{N(\omega)}\right\} \tag{28}$$

where $\mathcal{O}_N$ is the set of all images of objects that correspond to the lexical category under the noun N, $\mathcal{O}_A$ is the set of objects with the property consistent with adjective A, $\mathcal{U}_N$ is the set of all utterances of noun N, and $\mathcal{U}_{N(\cdot)}$ is the set of all utterances of the noun referring to object $\cdot$, similarly for adjective utterances. In testing the immature model by Eqs. (25) and (26), the model is presented with either a simultaneous visual appearance of an object and the utterance of its name, or the simultaneous visual appearance of the object and the utterance of its adjective. Tests of the mature model by Eqs. (27) and (28) require the simultaneous presentation of a visual object and its noun utterance, followed by the delayed adjective utterance. In the first case the class collects all objects and possible adjectives pertaining to a single noun, while in the second case the class collects all objects and names pertaining to a single adjective. As an alternative to Eq. (28), stimuli of the form $\langle o, n, a\rangle$ will be used to test ungrammatical sentences `[Noun Adj]`*.

Figure 4 show several cases of population coding for both nouns and adjectives, in the PFC and in the PFC* maps of the two models. In the case of nouns the spread and the amount of coding neurons is similar for the two models. The situation is different in the case of adjectives, the weakness of the coding in PFC compared to PFC* can be appreciated visually. For example in `yellow` and `blue` the amount

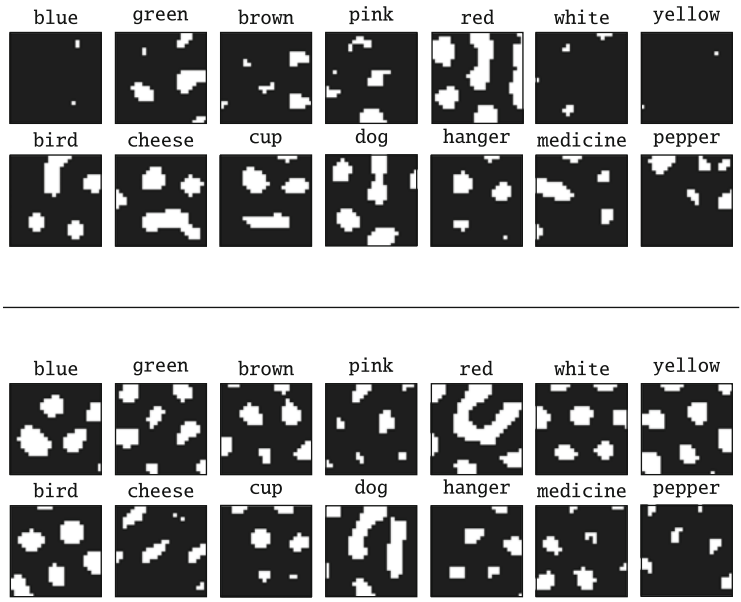

Fig. 4 Examples of population coding of objects and adjectives in the *upper map* of the models: PCF in the *top rows*, PFC* in the *bottom rows*

of coding neurons is tiny, and is very small also for `white`. In the PFC*, on the contrary, the amount and the distribution of the coding neuron is even for all adjectives.

3.3.1 Comparison of the Immature Model and the Model with Working Memory

In comparing the two versions of the model we introduce a metric for evaluating how accurate the knowledge acquired about noun and adjective lexical concepts is. The population code $p(S)$ computed with (24) can be used to classify a stimulus s in an expected category:

$$c(s) = \arg\max_{S \in \mathcal{S}} \left\{ \sum_{j=1}^{M} \alpha^j x_{p(S)_j}(s) \right\}, \tag{29}$$

where $p(S)_j$ denotes the j-th element in the ordered set $p(S)$ and α is a constant that is close, but smaller, than one. It is possible to evaluate how the population code in PFC and PFC* map is effective in discriminating a category S by measuring the fraction of hits in classifying stimuli belonging to that category:

$$a(S) = \frac{|\{s : s \in S \wedge c(s) = S\}|}{|S|}. \tag{30}$$

Table 2 Model accuracy in discriminating adjectives

Color	PFC	PFC* [Adj Noun]	[Noun Adj]*
Yellow	0.269	0.845	0.241
Red	0.518	0.789	0.743
Green	0.297	0.988	0.671
White	0.184	0.922	0.893
Brown	0.378	0.997	0.678
Pink	0.528	0.789	0.853
Blue	0.246	1.000	0.863
Mean	0.368 ± 0.19	0.903 ± 0.081	0.715 ± 0.226

The immature PFC achieved an accuracy on object nouns of 0.79 ± 0.25, the PFC* map of 0.52 ± 0.36, good levels, compared to the discrimination by chance of 0.026.

Table 2 displays the accuracy achieved at the end of development, comparing PFC, the upper map in the immature model, and PFC*, same map with working memory loop. In this case, adjectives are presented in both grammatical and ungrammatical sentences. Both versions of the model show the ability to achieve a good degree of recognition of color adjectives, largely above the chance threshold of 0.11. However, there is a significant improvement when working memory is in place, in all adjectives. In the less developed or immature model, accuracy is greater for nouns than for adjectives, which is rather contradictory, since there are 38 nouns as opposed to 9 adjectives, and noun categories easily cross boundaries of perceptual traits, confirming that it is the stage of the model that hampers adjective learning with respect to nouns.

It is interesting to note that when the sequence in the sentence is ungrammatical, the advantage in the comprehension of adjectives is reduced by half.

Therefore, the model in the PFC* version shows a syntact selectivity, in responding better to sentences where words respect their roles, however, this behavior is not in the form of a fixed rule, in that the violation of the syntax makes the adjective more difficult, but not impossible, to recognize.

3.4 Patterns of Connectivity of Nouns and Adjectives

All functions in the model arise spontaneously as a result of neural learning mechanisms, induced by exposure to stimuli. However, as long as cortical maps proceed in an anterior direction, the connectivity to sensorial input is more indirect and vague. An interesting investigation would be that of seeing whether representations in the higher model map of the two linguistic classes of color adjectives and object nouns differ in their patterns of connectivity to the lower map, which in the brain are more posterior and more directly related with stimuli.

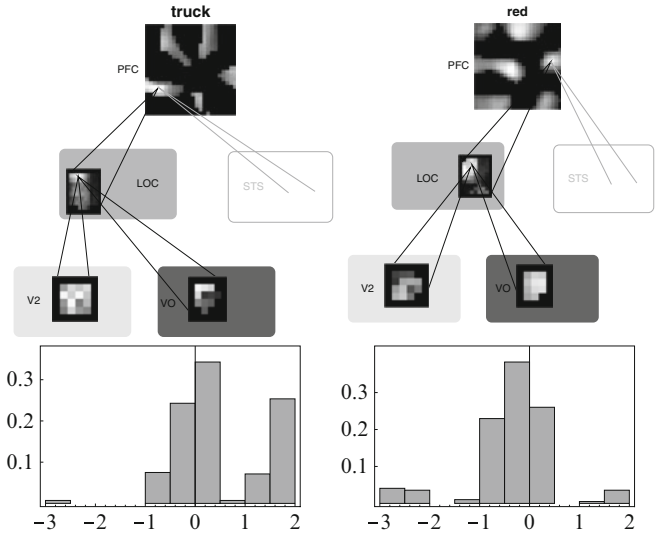

Fig. 5 The *top* two schemes are a comparison example of connectivity patterns for a noun (*left*) and an adjective (*right*). The *bottom* plots are the distribution of ξ connectivity parameter for all nouns (*left*) and all adjectives (*right*)

In particular, we analyzed differences in connectivity with respect to two visual processes: shape and color, which are segregated in the model in maps V2 and VO. For this purpose we define a parameter ξ_C, that measures the different amount of connections from the shape processing stream with respect to the color stream, for the population of neurons that code the category C. It is based on a preliminary parameter χ_C defined as:

$$\chi_C = \sum_{i \in \mathcal{X}_C^{(LOC)}} \frac{\Xi_\theta \left(\mathbf{a}_{r_{\mathrm{A}a,i}}^{(\mathrm{LOC} \leftarrow V2)} \right) - \Xi_\theta \left(\mathbf{a}_{r_{\mathrm{A}a,i}}^{(\mathrm{LOC} \leftarrow VO)} \right)}{\Xi_\theta \left(\mathbf{a}_{r_{\mathrm{A}a,i}}^{(\mathrm{LOC} \leftarrow V2)} \right) + \Xi_\theta \left(\mathbf{a}_{r_{\mathrm{A}a,i}}^{(\mathrm{LOC} \leftarrow VO)} \right)} \tag{31}$$

where receptive fields $\mathbf{a}$ are those of Eq. (14), the function $\Xi_\theta(\cdot)$ returns the number of connections in the receptive field $\cdot$ whose synaptic strength is larger than θ, and $\mathcal{X}_C^{(LOC)}$ is the set of neurons in LOC area projecting maximally into the population of neurons in PFC coding for category C. Now the parameter ξ is just the normalization of χ with respect to all the set of categories (both adjectives and nouns), to take into account a natural discrepancy in projections from V2 and VO due to the differences in size and architecture of the two maps:

$$\xi_C = \frac{\chi_C - \overline{\chi}}{\overline{\chi}} \tag{32}$$

Therefore positive values of ξ indicate a pattern of connectivity stronger towards shape processing areas, while negative towards color processing ones. Figure 5, in the bottom, shows the distribution of ξ traced back from the PFC population of neurons coding for all nouns, compared with the same distribution for all color adjectives. There is a significant difference in the distributions, in that color adjectives seem to recruit more from afferents coming from the color processing pathway than those coming from the shape processing pathway, compared to nouns. This can be interpreted as evidence of a physical grounding of the different meanings of the two linguistic classes, in the neural circuitry.

4 Conclusions

The model here described attempts to simulate lexical acquisition from auditory and visual stimuli from a brain processes point of view. The results of the modeling work show that cortical-like neural maps are able to detect and store coincidental associations in the stimuli, and build lexical categories: associations between words and visual concepts. Labels for objects, or nouns, are learned by a developmentally less mature version of the model very well, but adjectives, on the other hand, are not. In the more developmentally mature model, when working memory loops become available, adjectives become easier to learn. Furthermore, the more developed model, when presented with `[Noun Adj]`* sentences shows a decreased ability to recognize adjectives again, which we interpret as the model demonstrating an early sensitivity to a very basic form of syntax. Eventually, the backwards analysis of connections from the pre-frontal model map reveals an explanation of why some neurons form populations coding for nouns, and others for color adjectives. This explanation is to be found in the different recruiting of afferences from shape visual processing areas, or color processing areas.

References

1. Aboitiz, F., Garcia, R. R., Bosman, C., & Brunetti, E. (2006). Cortical memory mechanisms and language origins. *Brain and Language, 98*, 40–56.
2. Anzai, A., Peng, X., & Essen, D. C. V. (2007). Neurons in monkey visual area V2 encode combinations of orientations. *Nature Neuroscience, 10*, 1313–1321.
3. Ashby, F. G., & Spiering, B. J. (2004). The neurobiology of category learning. *Behavioral and Cognitive Neuroscience Reviews, 3*, 101–113.
4. Asr, F. T., Fazly, A., & Azimifar, Z. (2012). From cues to categories: A computational study of children's early word categorization (this volume), 81–104.
5. Atzori, M., Lei, S., Evans, D. I. P., Kanold, P. O., Phillips-Tansey, E., McIntyre, O., & McBain, C. J. (2001). Differential synaptic processing separates stationary from transient inputs to the auditory cortex. *Neural Networks, 4*, 1230–1237.
6. Baddeley, A. (1992). Working memory. *Science, 255*, 556–559.

7. Bates, E., Dal, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher, & B. M. Whinney (Eds.), *Handbook of child language* (pp. 96–151). Oxford, UK: Basil Blackwell.
8. Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Cognitive Brain Research, 403*, 309–312.
9. Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129–135.
10. Black, A. W., & Taylor, P. A. (1997). The festival speech synthesis system: System documentation (Tech. Rep. HCRC/TR-83). Human Communciation Research Centre, University of Edinburgh, Edinburgh, UK.
11. Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience, 12*, 3139–3161.
12. Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.
13. Booth, A. E., & Waxman, S. R. (2002). Word learning is smart: Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition, 84*, B11–B22.
14. Booth, A. E., & Waxman, S. R. (2008). Taking stock as theories of word learning take shape. *Developmental Science, 11*(2), 185–194.
15. Bornstein, M. H., & Cote, L. R. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, Dutch, French, Hebrew, Italian, Korean, and American English. *Child Development, 75*, 1115–1139.
16. Brown, M. C. (2003). Audition. In L. R. Squire, F. Bloom, S. McConnell, J. Roberts, N. Spitzer, & M. Zigmond (Eds.), *Fundamental neuroscience* (pp. 699–726). New York: Academic.
17. Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). Cambridge, MA: MIT Press.
18. Carey, S. (2009). *The origin of concepts*. Oxford, UK: Oxford University Press.
19. Clark, E. V. (1973). What's in a word: On the child's acquisition of semantics in his first language. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic.
20. Dannemiller, J. L. (1989). A test of color constancy in 9- and 20-weeks-old human infants following simulated illuminant changes. *Developmental Psychology, 25*, 171–184.
21. Darwin, C. (1877). A biographical sketch of a young infant. *Kosmos, 1*, 367–376.
22. Dowling, J. E. (1987). *The retina: An approachable part of the brain*. Cambridge, UK: Cambridge University Press.
23. Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci2008)* (pp. 703–708). Austin, TX: Cognitive Science Society.
24. Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology, 88*, 929–941.
25. Fuster, J. M. (2001). The prefrontal cortex–an update: Time is of the essence. *Neuron, 30*, 319–333.
26. Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language and Cognitive Processes, 13*, 269–306.
27. Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. *Cognitive Development, 49*, 988–998.
28. Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. *Child Development, 75*, 1098–1114.
29. Goldman-Rakic, P. S. (1995). Architecture of the prefrontal cortex and the central executive. *Annals of New York Academy of Science, 769*, 71–83.
30. Grassman, S., Stracke, M., & Tomasello, M. (2009). Two year olds exclude novel objects as potential referents of novel words based on pragmatics. *Cognition, 112*, 488–493.

31. Grill-Spector, K., Kushnir, T., Edelman, S., Avidan-Carmel, G., Itzchak, Y., & Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron, 24*, 187–203.
32. Hall, G., & Waxman, S. R. (Eds.). (2004). *Weaving a lexicon*. Cambridge, MA: MIT Press.
33. Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*, 393–402.
34. Hirsh-Pasek, K., Golinkoff, R. M., Hennon, E. A., & Maguire, M. J. (2004). Hybrid theories at the frontier of developmental psychology: The emergentist coalition model of word learning as a case in point. In G. Hall, & S. R. Waxman (Eds.), *Weaving a lexicon* (pp. 173–204). Cambridge, MA: MIT Press.
35. Hollich, G., Golinkoff, R., & Hirsh-Pasek, K. (2007). Young children associate novel words with complex objects rather than salient parts. *Developmental Psychology, 43*, 1051–1061.
36. Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). *Breaking the language barrier: An emergentist coalition model for the origins of word learning* (Vol. 65). Oxford, UK: Basil Blackwell. Monographs of the Society for Research in Child Development.
37. Hubel, D. H., & Wiesel, T. N. (1959). *Brain and visual perception: The story of a 25-year collaboration*. Oxford, UK: Oxford University Press.
38. Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.
39. Huey, E. D., Krueger, F., & Grafman, J. (2006). Representations in the human prefrontal cortex. *Current Directions in Psychological Science, 15*, 167–171.
40. Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *Journal of Neuroscience, 24*, 3313–3324.
41. Katz, L., & Shatz, C. (1996). Synaptic activity and the construction of cortical circuits. *Science, 274*, 1133–1138.
42. Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology, 71*, 856–867.
43. Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.
44. Kourtzi, Z., Erb, M., Grodd, W., & Bülthoff, H. H. (2003). Representation of the perceived 3-d object shape in the human lateral occipital complex. *Cerebral Cortex, 13*, 911–920.
45. Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language, 31*, 807–825.
46. Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15*, 1621–1631.
47. Linden, J. F., & Schreiner, C. E. (2006). Columnar transformations in auditory cortex? A comparison to visual and somatosensory cortices. *Cerebral Cortex, 13*, 83–89.
48. Macnamara, J. (1982). *Names for things: A study of human learning*. Cambridge, MA: MIT Press.
49. Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 57–77.
50. Mastronarde, D. N. (1983). Correlated firing of retinal ganglion cells: I. spontaneously active inputs in X- and Y-cells. *Journal of Neuroscience, 14*, 409–441.
51. Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review, 117*, 1–31.
52. Miller, E. K., Freedman, D. J., & Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions: Biological Sciences, 357*, 1123–1136.
53. Miller, L. M., Escab, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *Journal of Neurophysiology, 87*, 516–527.
54. Mintz, T. H., & Gleitman, L. R. (2002). Adjectives really do modify nouns: the incremental and restricted nature of early adjective acquisition. *Cognition, 84*, 267–293.
55. Mountcastle, V. (1957). Modality and topographic properties of single neurons in cats somatic sensory cortex. *Journal of Neurophysiology, 20*, 408–434.

56. Nayar, S., & Murase, H. (1995). Visual learning and recognition of 3-d object by appearence. *International Journal of Computer Vision, 14*, 5–24.
57. Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science, 6*, 136–142.
58. Plebe, A. (2007). A model of angle selectivity development in visual area V2. *Neurocomputing, 70*, 2060–2066.
59. Plebe, A. (2008). The ventral visual path: Moving beyond V1 with computational models. In T. A. Portocello, & R. B. Velloti (Eds.), *Visual cortex: New research* (pp. 97–160). New York: Nova Science Publishers.
60. Plebe, A., & Domenella, R. G. (2006). Early development of visual recognition. *BioSystems, 86*, 63–74.
61. Plebe, A., & Domenella, R. G. (2007). Object recognition by artificial cortical maps. *Neural Networks, 20*, 763–780.
62. Plebe, A., Mazzone, M., & De La Cruz, V. M. (2010). First words learning: A cortical model. *Cognitive Computation, 2*, 217–229.
63. Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development, 77*, 266–280.
64. Quine, W. V. O. (1960). *Word and object*. New York: Columbia University Press.
65. Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy, 11*, 232–253.
66. Rogers, T. T., & McClelland, J. L. (2006). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
67. Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*, 113–146.
68. Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.
69. Samuelson, L. (2002). Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20-month-olds. *Developmental Psychology, 38*, 1016–1037.
70. Sandhofer, C. M., & Smith, L. B. (2001). Why children learn color and size words so differently: Evidence from adults' learning of artificial terms. *Journal of Experimental Psychology, 130*, 600–620.
71. Schiller, P. H. (1996). On the specificity of neurons and visual areas. *Behavioural Brain Research, 76*, 21–35.
72. Sirosh, J., & Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation, 9*, 577–594.
73. Sloutsky, V. M., & Robinson, C. W. (2008). Flexible attentional learning in infancy. In V. M. Sloutsky, B. C. Love, & K. McRae (Eds.), *Proceedings of the XXX Annual Conference of the Cognitive Science Society* (pp. 1182–1187). Mahwah, NJ: Lawrence Erlbaum Associates.
74. Smith, L. B. (1999). Children's noun learning: How general learning processes make specialized learning mechanisms. In B. MacWhinney (Ed.), *The emergence of language* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
75. Swingley, D. (2010). Fast mapping and slow mapping in children's word learning. *Language Learning and Development, 6*, 179–183.
76. Thompson, I. (1997). Cortical development: A role for spontaneous activity? *Current Biology, 7*, 324–326.
77. Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
78. Ungerleider, L., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of visual behavior* (pp. 549–586). Cambridge, MA: MIT Press.

79. Van Essen, D. C., & DeYoe, E. A. (1994). Concurrent processing in the primate visual cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences.* Cambridge, MA: MIT Press.
80. Vanduffel, W., Tootell, R. B., Schoups, A. A., & Orban, G. A. (2002). The organization of orientation selectivity throughout the macaque visual cortex. *Cerebral Cortex, 12*, 647–662.
81. Verkindt, C., Bertrand, O., Echallier, F., & Pernier, J. (1995). Tonotopic organization of the human auditory cortex: N100 topography and multiple dipole model analysis. *Electroencephalography and Clinical Neurophisiology, 96*, 143–156.
82. Wandell, B. A. (1999). Computational neuroimaging of human visual cortex. *Annual Review of Neuroscience, 10*, 145–173.
83. Wandell, B. A., Brewer, A. A., & Dougher, R. F. (2005). Visual field map clusters in human cortex. *Philosophical Transactions of the Royal Society of London*, *360*, 693–707.
84. Weigelt, S., Kourtzi, Z., Kohler, A., Singer, W., & Muckli, L. (2007). The cortical representation of objects rotating in depth. *Journal of Neuroscience, 27*, 3864–3874.
85. Winer, J. A., Miller, L. M., Lee, C. C., & Schreiner, C. E. (2005). Auditory thalamocortical transformation: Structure and function. *Neuron, 28*, 255–263.
86. Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience, 4*, 139–147.
87. Zeki, S. (1983). Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelenghts and colours. *Neuroscience, 9*, 741–765.
88. Zeki, S. (1983). Colour coding in the cerebral cortex: The responses of wavelength-selective and colour-coded cells in monkey visual cortex to changes in wavelength composition. *Neuroscience, 9*, 767–781.

Part III
Learning Morphology and Syntax

Treebank Parsing and Knowledge of Language

Sandiway Fong, Igor Malioutov, Beracah Yankama, and Robert C. Berwick

Abstract Over the past 15 years, there has great success in using linguistically annotated sentence collections, such as the Penn Treebank (PTB), to construct statistically based parsers. This success leads naturally to the question of the extent to which such systems acquire full "knowledge of language" in a conventional linguistic sense. This chapter addresses this question. It assesses the knowledge attained by several current statistically-trained parsers in the area of tense marking, questions, English passives, and the acquisition of "unnatural" language constructions, extending previous results that boosting training data via targeted examples can, in certain cases, improve performance, but also indicating that such systems may be too powerful, in the sense that they can learn "unnatural" language patterns. Going beyond this, this chapter advances a general approach to incorporate linguistic knowledge by means of "linguistic regularization" to canonicalize predicate-argument structure, and so improve statistical training and parser performance.

1 Introduction: Treebank Parsing and Knowledge of Language

Parsers statistically trained on corpora like the Wall Street Journal/Penn Tree Bank have steadily improved their performance. However, despite these gains, it is well-known that such systems often perform poorly on novel sentences

S. Fong (✉)
University of Arizona, Tuscon, AZ 85721, USA
e-mail: sandiway@email.arizona.edu

I. Malioutov · B. Yankama · R.C. Berwick
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: igorm@mit.edu; beracah@.mit.edu; berwick@csail.mit.edu

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_6,

outside their training datasets, due to the sparsity effects that reflect the "long-tail" Zipf-distributional rarity of linguistic constructions and head-dependency relations (see Collins [15], among many others). Klein and Manning [27] summarize the situation in this way:

> As a speech person would say, one million words of training data just isn't enough. Even for topics central to the treebank's WSJ text, such as stocks, many very plausible dependencies occur only once, for example, *stocks stabilized*, while many others occur not at all, for example, *stocks skyrocketed*.

Our experiments below suggest that sufficiently complex linguistic constructions exhibiting non-local dependencies may often pose problems for a parsing model that takes a static view of syntactic structure – a model unable to systematically relate the passive form of a sentence to its active counterpart, or a declarative sentence to a corresponding derived interrogative. While often effective, simply adding more data should not be invariably seen as a substitute for incorporating explicit linguistic constraints into parsing models. Indeed, the successful use of an alternative model of syntactic structure, Combinatory Categorial Grammar (CCG), as implemented in several recent systems such as the C&C parser [11] and by Hockenmaier [22,23] may be seen as a concrete demonstration that sometimes the representation of syntactic knowledge, rather than data sparsity, plays a more important role in parser performance.

Moreover, as evidenced by the Penn Treebank, more challenging linguistic mechanisms may have the least amount of data available for learning. The problem is only exacerbated if we examine resource-impoverished languages. Language acquisition is a classic instance of a scenario where adding more data is not one of the available options for resolving the data sparsity problem. A viable computational treatment requires model-level changes to address this issue.[1]

In fact, our experiments below indicate that statistical parsing stands to benefit from a much more restrictive learning regime that inherits insights from language acquisition. On this view, parsing models should be judged based on their ability to recover and discriminate between different types of syntactic mechanisms rather than on incremental improvements from adding training data to alleviate the data sparsity problem. Similarly, the ability of a model to learn an unnatural syntactic mechanism detracts from its ability to discriminate between syntactic constraints observable in human language. Conversely, insights from our experiments can be

[1] We note that there have been recent proposals that suggest that "linguistic mastery does not need to be available early in the course of language development" and that "the acquisition of usage-based and fixed-form patterns can account for ... [the] syntactic burst [occuring around age two to three]" [39]. It is uncontroversial that some fixed form patterns are memorized by children, and equally that complete linguistic mastery of syntax is delayed until the age of eight or later, as first established by the work of Carol Chomsky [10]. However, while it "need not" be "available early", in point of fact, empirically, it has long been established that 'telegraphic speech' is not indicative of the full scope of syntactic comprehension at the ages of 2–3; rather, many aspects of syntax are acquired by this age, but telegraphic speech does not reveal these abilities and reveals processing difficulties such as memory limitations [20,47].

brought to bear on approaches to language acquisition. Syntactic mechanisms might be more effectively acquired and discriminated if they are characterized in terms of canonical argument analysis.

More generally, in this chapter we will focus on an assessment of gaps in the "knowledge of language" acquired by statistically-trained parsers, attempting to sort out which of these might arise from limited training data and lead to parameter estimation problems with associated parsing models, and which might arise from underlying grammatical frameworks and benefit from the insights of linguistic theory.

We note that often the two sources of error are not complementary. Adding more data relevant to a particular syntactic construction may resolve parsing mistakes, but at the same time it may be symptomatic of a systematic problem with the model. When asked to chose between two solutions, their relative ability to scale up and generalize to new instances is the critical consideration. For example, a model that needs a passive form for each active counterpart observed in the data to be able to parse the passive variant should be less preferred to a model that explicitly models the passive and is able to analyze and generate such a form automatically. This is the basic conclusion we draw from our analysis of passive sentences, and it is not simply a question about data sparsity.

We should emphasize at the outset that we have probed questions like these by constructing entirely new experiments, not simply covering familiar ground about the ever-present issue of data sparsity in statistical parsing. To the best of our knowledge, all our experiments and their results are new. The analysis of passive errors and the method we apply to canonicalize argument structure to improve passive parsing performance is also novel, as far as we have been able to determine. Similarly, our analysis of wh-questions does not simply rehash the approach of Rimmell et al. [44]. Finally, our application of an "unnatural" language learning litmus tests, while drawn from the psycholinguistic literature as in [36], has not been extended to current statistical parsers. In all of these situations, our ultimate goal is to seek ways of improving parsers by determining whether such systems have typical failure modes that can be discovered, as well as whether these failures need to be remedied.

To begin, such an assessment of "knowledge of language" poses a real challenge. Parsers are typically designed from the start to solve a very particular engineering task that is quite different from the way that a linguist might assess knowledge of language. Roughly speaking, statistically-based parsers learn how to select a "most likely" analysis with respect to all the parses they have been trained on and all the parses they can generate. They only choose among possible parses, standardly using either generative or discriminative estimation methods. In this sense, they do not directly adjudicate among "grammatical" and "ungrammatical" sentences.[2] Such a

[2]As noted in [41] and [48], despite the fact that statistically-based parsers have used both sorts of estimation methods, the underlying statistical models for both generative approaches as well as discriminative approaches using what are called "latent variables" – probabilistic and weighted context-free grammars, respectively – turn out to be equivalent in their expressive power.

probabilistic "remembrance of parses past" is not the same as the replicability of linguistic knowledge conventionally probed by grammaticality judgements.

Indeed, it is not immediately obvious how to align grammaticality judgements with probabilities. There is no agreed-upon unification. While some authors, e.g., Abney [1] maintain that the grammaticality-probability distinction should be kept firmly apart, still others argue differently, e.g., [29], p. 33:

> The parser that an ML [machine learning] system produces can be engineered as a classifier to distinguish grammatical and ungrammatical strings.

While a more detailed consideration of this point lies beyond the scope of this chapter, it suffices to observe that, as noted in [12], one cannot simply provide a probability threshold, ϵ, such that for all probability values greater than ϵ, a parse is grammatical, otherwise ungrammatical. In this case there could be at most $1/\epsilon$ grammatical sentences, and the corresponding language would be finite. Observe that the standard assumption for probabilistic context-free grammars assumes an exponential distribution of probability mass with respect to generated sentence length, so that sentences longer than a certain length have vanishingly small probability mass. Such a language is effectively finite. If anything, to the extent that such parsers are intended to model an actual corpus, they presumably reflect actual language *use*, (in the case of the PTB, newspaper writing), and so a complex mix of syntactic, lexical-semantic, world/encyclopedic knowledge, processing load, and other similar factors. This is not coextensive with the conventionally abstract, linguistic notion of linguistic *competence*, that deliberately idealizes away from this mix, though there are familiar points of contact.

Consequently, in this chapter we will typically base our assessments simply on what parsing systems can and cannot do well. To consider an introductory example of the assessment methods we will use, even in simple cases many corpus-trained parsing systems cannot recover correct verb argument structure. Consider a passive construction such as that in Ex. 1 below:

(1) Mary was kissed by the guy with a telescope on the lips.

Many (perhaps most) parsers trained on the PTB will tend to attach the Prepositional Phrase (PP) *on the lips* incorrectly to the PP *a telescope* because most of their training data follow such a form. In contrast, the corresponding active form, Ex. 2 below, is easily parsed correctly by such systems, because the Subject NP-PP combination is no longer located near the ambiguous PP attachment point:

(2) The guy with a telescope kissed Mary on the lips.

Such examples are not just hypothetical. For instance, Fig. 1 shows that sentence #404 of section 23 of the PTB, *Measuring cups may soon be replaced by tablespoons in the laundry room*, is parsed incorrectly exactly in this way by two state-of-the-art parsers, the Stanford unlexicalized context-free parser [27] and Bikel's re-implementation of the Collins parser [4]. In all these cases, the PP *in the laundry room* is incorrectly attached as a modifier of the object NP *tablespoons*.

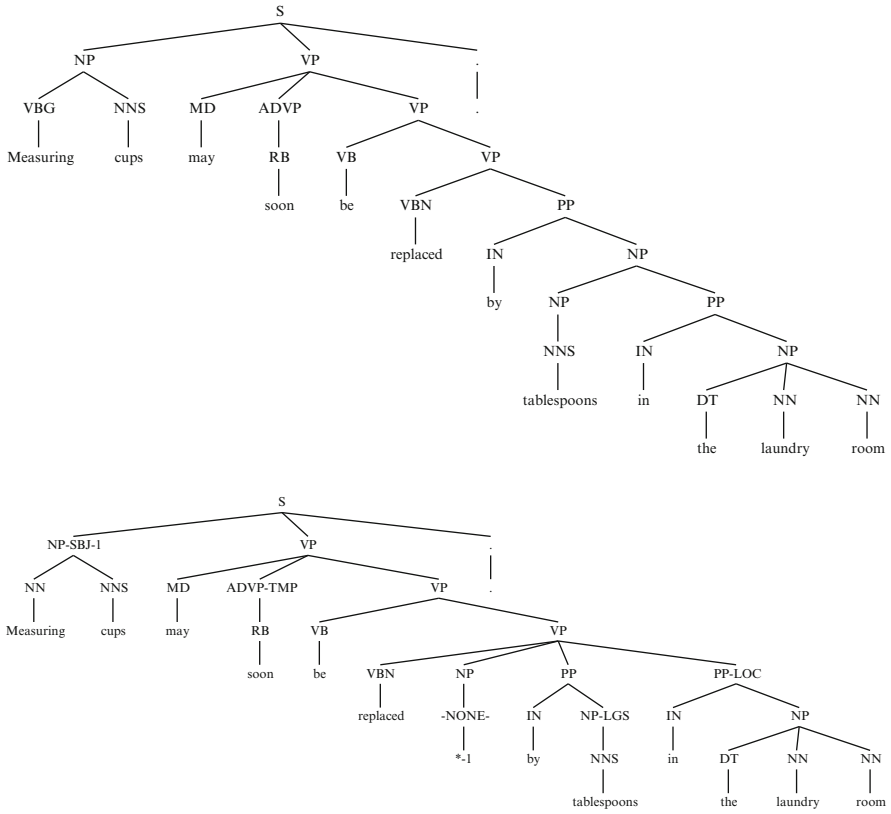

Fig. 1 The Bikel/Collins and Stanford unlexicalized parsers both mis-analyze sentence number 404 in section 23 of the PTB. The *top* half of the figure shows the result of parsing using either Bikel's reimplementation of the Collins parser or the Stanford unlexicalized parser. The *bottom* half of the figure shows the corresponding "gold standard" PTB structure

In the remainder of this chapter, with some exceptions we will typically test examples on a range of probabilistic parsers in an attempt to avoid the idiosyncrasies of any particular implementation and achieve some measure of robustness in our test results. In this case, in addition to the two parsers illustrated in the main text, the Berkeley parser [40] and the C&C combinatory categorial grammar parser [18] both output the same, incorrect attachment. The Malt dependency parser version 1.4.1 [37] also outputs an incorrect dependency between *in* and *tablespoons*. In contrast, both the "factored" Stanford lexicalized-dependency parser [28] and the Charniak-Johnson parser [6] *do* output the correct attachment.

Examples such as these suggest that verb argument structure might be more easily recoverable when sentence structure is represented in some canonical format that more transparently encodes grammatical relations such as Subject and Object. In other words, if the arguments of predicates are in a fixed syntactic position

in training examples, then we might expect that this regularity would be simpler for a statistically-based system to detect and acquire. More generally, it has often been observed that what makes natural languages difficult to acquire or parse is that phrases are displaced from their canonical positions, not only in passives, but in topicalization, wh-movement, and many similar constructions. Each of these constructions breaks the transparent link between predicates and arguments. In Sect. 5 below, we shall see that one can remedy at least some of these difficulties by adopting a representation that is arguably closer to the one that certain linguistic theories assume, where the argument of the main verb has been 'replaced' in its canonical Object position, as in Ex. 1. There are other representations one might adopt to handle this particular problem, for example, a combinatorial categorial grammar (CCG) that explicitly relates displaced phrases to their "gaps." As we noted earlier, this does not necessarily ensure success.

Following the lead of this illustrative example, in the remainder of this chapter we will focus on the following selection of challenging areas for parsers trained on corpuses like the PTB:

1. **Wh-questions.** As has often been noted, the PTB corpus contains a very small number of questions – unsurprisingly, since it consists of *Wall Street Journal* newspaper articles [34]. Out of the 39,822 sentences in the standard training sections 02–21, there are only 128 "root" level questions, such as training data sentence #85, *What's next?* and four other similar questions. More than 70 % of these are Subject wh-questions There are 61 additional wh-questions that appear in embedded quotational contexts, e.g., *"What's he doing " , hissed my companion*, and 96 root level auxiliary inverted questions, e.g., *Was this why some of the audience departed before or during the second half* . In short, by all measures, the training data for wh-constructions and questions is exceptionally sparse. Moreover, the statistically-trained parsers we examine in this chapter do not receive data in the form of "more ill-formed" examples that differ, say, by just a single word in a different order, such as, *Who asked who bought what* vs. *Who asked what who bought*. These systems must therefore learn such nuances from just one or two positive examples.
2. **Tense marking.** Tense is a good example of a linguistic phenomenon that, like displacement in wh-questions, may be "spread out" over several, not necessarily adjacent words. For example, in an English yes-no question, tense must be realized overtly at the front, while the corresponding main verb need not have an overt morphological indicator of tense: thus we have the PTB example, *Do you think the British know something we don't* , where *do* carries tense and *think* does not. We will investigate whether statistically-trained systems can "capture" part of the English tense system by examining examples of verbs that are ambiguously marked for tense, such as *read* or *cost*.
3. **Passives.** As noted in our introductory example, the placement of a verb's argument in Subject position, along with the possibility of an Agentive "by" phrase can lead to parsing difficulties.

4. **"Unnatural" language constructions.** Finally, while the previous topics all examine a particular parsing task – essentially, structural language patterns – that one would like a trained parser to detect easily, there are also non-attested language patterns that trained parsers should be able to detect only with great *difficulty*. A cognitive-faithful parser should have the same problems acquiring "unnatural" language patterns as people do. But what do mean by unnatural? By this we do not mean patterns that are challenging for people due to processing constraints, e.g., the classic examples of center-embedded or garden path constructions. Rather, what we will mean by "unnatural" language constructions are examples of the sort studied in some detail by Musso et al. [36] via artificial grammar learning and fMRI experiments. They covered two sorts of unnatural rules: (1) "counting" rules, that is, linguistic rules that, say, could form the negation of a declarative sentence by inserting a special word at a particular point in a sentence, say, always immediately after the third word; (2) "mirror image" rules, that is, linguistic rules that, say, could form the interrogative of a declarative sentence by inverting the word order of the declarative sentence, saying it in reverse. In their study, [36] constructed a set of unnatural rules, unattested in any natural language. Here is their description of the second "unnatural" rule, which is the one in Sect. 6 that we will attempt to reproduce as closely as possible in our experiments with statistical parsers, from [36], p. 775:

   > The second rule required that the interrogative construction be built by inverting the linear sequence of words of a sentence. For example, "*I* [1] *bambini* [2] *amano* [3] *il* [4] *gelato* [5] or "The children love ice-cream" becomes *Gelato* [5] *il* [4] *amano* [3] *bambini* [2] *il* [1].

   Musso et al. found that people had great difficulty mastering artificial rule systems of this sort. If they were learned at all, they were learned, as if they were non-linguistic 'puzzles,' activating very different brain regions than those lit up during normal language rule processing. Smith et al. [49] reported a similar finding, again using an artificial grammar learning paradigm. Here it was discovered that an autistic linguistic "savant" could not learn "unnatural" grammatical rules. In contrast, while adults could learn these rules, but again, only with great difficulty. In a related area, others (e.g., [33]) have noted that the same issue arises with respect to artificial neural network learning in the paradigm case of English past tense over-regularization. Neural network systems that are constructed to report the probability of the next word or form in a sequence are apparently "unnatural" to the extent that they can learn sentence reversals just as easily as normally ordered word sequences. Note that this is a case where the neural network simulations do equate "grammaticality" with "likelihood." What all these results come to is the same: we do not want a "natural" learning system to be *too* flexible, having capacities beyond those found in people.[3]

[3] See, e.g., [9] and [2] for additional discussion of the lack of non-counting and palindromic rules in natural language, including syntax and phonology. It is known in certain sociological settings

2 Experimental Methods

We carried out our experiments on as broad a range of publicly available statistically-trained parsers as possible, subject to the broad constraint they all could be trained on the same, standard subsections of the *Wall Street Journal* version of the Penn Tree Bank III. In this we strove to follow the same procedure and roughly the same coverage as in the comparative study carried out in [13], p. 51:

> Constituent parsers and dependency parsers all have the appropriate level of sophistication, but a wide variety of different grammars and conceptual frameworks that makes comparing them difficult. However, there is one class of parsers that is both numerous and up-to-date, and covers a variety of different algorithms which all use the same output format (bar a few small details). These are sometimes referred to as treebank parsers as they are usually trained and optimized on the PTB and produce output conformant with its standards.

2.1 Parsing Systems Used

The systems that were used for the experiments are given in Table 1. Not all of these systems could be used for all experiments, due to certain resource requirements. Such details will be noted in what follows. Among the publicly available systems, we selected the most extensively cited and most widely used parsers. We cannot hope to exhaust the full range of parsers now publicly available, particularly dependency parsers. For example, we could not include the Melamed/Turian discriminative parser [52]. We leave such extensions for future research. Additional details about the grammatical models and the training/testing procedures used will be covered as they arise.

2.2 Training Data, Testing, and Evaluation

In order to ensure that results would be as comparable as possible, we retrained most of the parsers on sections 02–21 of the PTB III, even when they came with "pre-built" estimated models on this training data (as with the C-J, Berkeley, and Stanford parsers).[4] Due to limited access to the original materials and other computational constraints, we were not able retrain the CJ-R parser. As a result, in what follows we

---

that palindromic forms are used, e.g., the Australian butchers' market language. But all indications here are that this such behavior remains "puzzle based."

[4]We attempted to use training settings that matched those for the parsers' "pre-built" models as far possible. For example, we used the settings provided in the Stanford parser directory under `makeSerialized.csh` for the so-called `wsjPCFG` model. In the case of the BC-M2 parser, we used the settings given by `collins.properties` since we wanted to ensure replicability with standard results.

Table 1 The treebank parsers chosen for this investigation

Parser	Abbreviation	Release used	Citation
Bikel-Collins Model 2	BC-M2	1.2 Oct 08[a]	[4]
Berkeley "coarse to fine"	Berkeley	1.1, Sept 09[b]	[40]
Stanford unlexicalized	Stanford-unlex	1.6.3[c]	[27]
Stanford factored dependency	Stanford-fact	1.6.3[c]	[28]
Charniak "coarse-to-fine"	CJ-I	Nov 09[d]	[5]
Charniak-Johnson reranking	CJ-R	Nov 09[d]	[6]

[a] http://www.cis.upenn.edu/~dbikel/download/dbparser/1.2/install.sh
[b] http://code.google.com/p/berkeleyparser/downloads/detail?name=berkeleyParser.jar
[c] http://nlp.stanford.edu/software/stanford-parser-2010-07-09.tgz
[d] http://web.science.mq.edu.au/~mjohnson/code/reranking-parser-Nov2009.tgz

used only the CJ-R pre-built model. In addition to using this standard training data, we carried out various experimental manipulations followed by data augmentation and retraining that will be described in later sections. For evaluation we used the standardly available `evalb` package [46].

3 Case Study: Parsing Wh-Questions and QuestionBank

We first return to the area of wh-questions outlined briefly in Sect. 1. For the purposes of this chapter, we will put to one side the question of how to link wh-words and phrase such as *what* or *which problem* to their 'gaps', for example, the link between *what* and the object position after *buy* in a sentence such as *What did John buy*. While this is an important topic, full analysis of this problem is beyond the scope of the current chapter; see [44] and [18] for combinatory categorial grammar approaches that address this issue. Instead we will focus solely on the question of how well correct parses are recovered.

Why would parsing problems arise even if we put this issue aside? The reason is that in the standard training sections of the PTB, wh-phrases are most often used as relative clauses, not as questions (in a ratio of approximately 10,000:1). It would not be surprising, then, if a true wh-question was parsed as if it were a relative clause. Using standard PTB notation, we would then expect wh-questions parsed incorrectly as an S embedded within an SBAR, rather than, correctly, as an SQ (a sentential question) embedded within an SBARQ. (See Fig. 2 below for a representative example of this distinction.)

To be concrete, a conventional linguistic assessment about knowledge regarding wh-questions often begins with a "graded" list of examples such as those in Ex. 3 below, where the first sentence is an "echo question." This is followed by a semantically similar wh-interrogative sentence. The next three examples are then listed in roughly an order of descending acceptability to native English speakers (hence the asterisks placed before them).

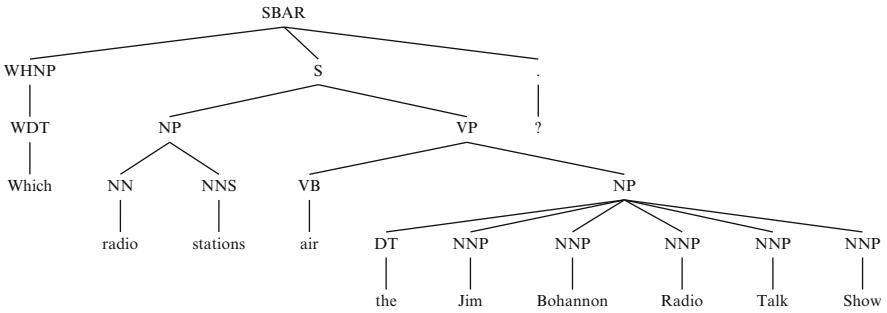

Fig. 2 An example of a wh-question parsing error for the sentence, *Which radio stations air the Jim Bohannon Radio Talk Show?* This is the output from the BC-M2 parser

(3) a. Bill will solve which problem?
b. Which problem will Bill solve?
c. Which problem Bill will solve?
d. Bill solve which will problem?
e. Which problem Bill solve will?

How might we use such examples to test the linguistic knowledge acquired by a statistically-trained parser? Note that even if a sentence is "ill-formed" like the last three above, then a probabilistic parser will still try to do the best it can, and return the most likely analysis, even a partial or incorrect one, with respect to the parsed examples it has already been trained on. That is in some respects an appropriate response to what such systems have been designed to do, one means to add robustness. As we described in the introduction, this might be a perfectly valid way to proceed from an engineering standpoint; factoring in gradience judgements of this sort remains an area to explore that lies beyond the scope of the present chapter. Further, while we might expect that the probability scores returned by the parser for the last three sentences could be worse than those for the first two, likelihood scores would probably vary anyway given slightly different local contexts and the successive history of various local rule choices set against what has been seen in the training corpus. In addition, if a parser is "lexicalized" then the actual word information (e.g., whether the verb is *solve* or *try*) is typically propagated to the head of a phrase (in this case, the Verb Phrase (VP)), and in this way specific lexical items may play a role in influencing what analysis path is taken.

Putting this question of assessing grammaticality to one side, we therefore focus instead only on the problem of producing the correct parse, rather than any likelihood score that denotes relative acceptability or grammaticality. That this is a real problem may be seen in Fig. 2 below, which displays an incorrect parse of a wh-question sentence produced by the BC-M2 parser, on an example sentence taken from an actual corpus of wh-questions, QuestionBank, that we describe immediately below.

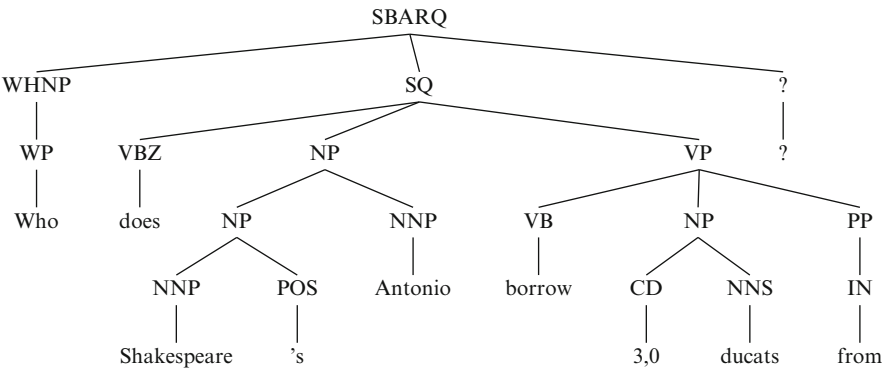

Fig. 3 Parse structure assigned to the "Who does Shakespeare. . ." sentence by the downloaded QuestionBank used in the current analysis

3.1 Augmenting the Training Data

There have been several approaches to remedying this problem by adding additional wh-question training sentences. In particular, Judge et al. [26], Rimmell et al. [44], and Nivre et al. [38] have built systematic "unbounded dependency" question treebanks.

We did not have access to these last resources, so we drew instead on a recently-built publicly accessible 4,000 sentence database, QuestionBank, constructed by Judge et al. [26]. This is a curated database of 2,000 questions drawn from the TREC question-answering (QA) domain and 2,000 questions from the Cognitive Computation Group at UUIC.[5] A representative example from this version of the QuestionBank is, *Who does Shakespeare's Antonio borrow 3,0 ducats from?*, as displayed in Fig. 3. Note that unlike the PTB II/III, this downloaded version did not contain information about the location of the underlying argument positions of displaced phrases, e.g., that *Who* serves as the object argument *from*) in the preceding example. From our perspective this was satisfactory because, unlike the research reported on in [26,44], or [38], we were interested solely in the question of whether statistical parsers could learn correct structural analyses.

Note that while QuestionBank represents approximately a 10 % addition to the number of sentences to the baseline training set, most of these wh-question sentences are typically far shorter than those in the PTB II, with a median sentence length of ten words – unsurprising since these are questions culled from a question-answering domain as opposed to the written *Wall Street Journal* newspaper article domain.

[5]The full database was obtained by download from http://www.computing.dcu.ie/~jjudge/qtreebank/. A handful of errors in corpus annotation were corrected in this downloaded dataset.

Table 2 Labeled precision, labeled recall, and F-Scores for baseline and wh-trained parsers, using question training/test data from QuestionBank (QB). The last column displays F-scores for these parsers' performance on only the standard baseline section 23 of the WSJ

Parser type	Labeled precision, %	Labeled recall, %	F-score, %	F-score, % WSJ Sect. 23
BC-M2 baseline	80.87	71.25	75.76	85.63
BC-M2+QB	91.08	81.70	86.18	85.79
% improvement	12.63	14.67	13.75	
Stanford-unlex baseline	66.26	69.32	67.57	85.54
Stanford-unlex+QB	81.72	80.92	81.32	85.55
% improvement	22.33	22.01	20.03	
Stanford-fact baseline	62.50	65.57	64.00	88.71
Stanford-fact baseline + QB	88.71	87.41	88.06	88.59
% improvement	20.53	15.60	17.99	
CJ-I baseline	84.65	71.81	77.70	86.55
CJ-I+ QB	90.31	80.65	85.21	88.13
% improvement	6.69	12.31	9.67	

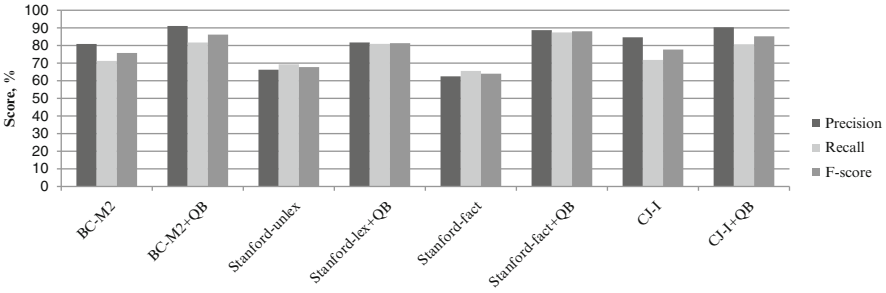

Fig. 4 Labeled precision, labeled recall, and F-scores for the parsers trained and tested on the QuestionBank corpus, both before and after training on QuestionBank

We divided the 4,000 QuestionBank sentences into an 80 % training portion and a 20 % testing portion. We tested four parsers: BC-M2; Stanford-lex; Stanford-fact; and CJ-I. We tested each of these four parsers on two training-test sets: (1) the baseline conventional PTB training set; (2) the 80 % Question Bank sample, eight experiments in all.

Table 2 gives the complete numerical results of these eight runs, while Fig. 4 displays the results visually, as histograms of the precision, recall, and F-score before/after performance. Both reveal a substantial improvement across all parsers. For example, Stanford-unlex parser had labeled precision/labeled recall scores of 66.26 %/69.32 % before training, and 81.72 %/80.92 % after training, a considerable gain of 15 and 10 % points, respectively (a 20.53 % and 15.60 % increase). The CJ-I parser's scores were boosted from 84.65 %/71.81 % to 90.31 %/80.65 % This was the smallest percentage improvement, due probably to the fact that even before wh-training the CJ-I parser already performed quite well. Still, increases with

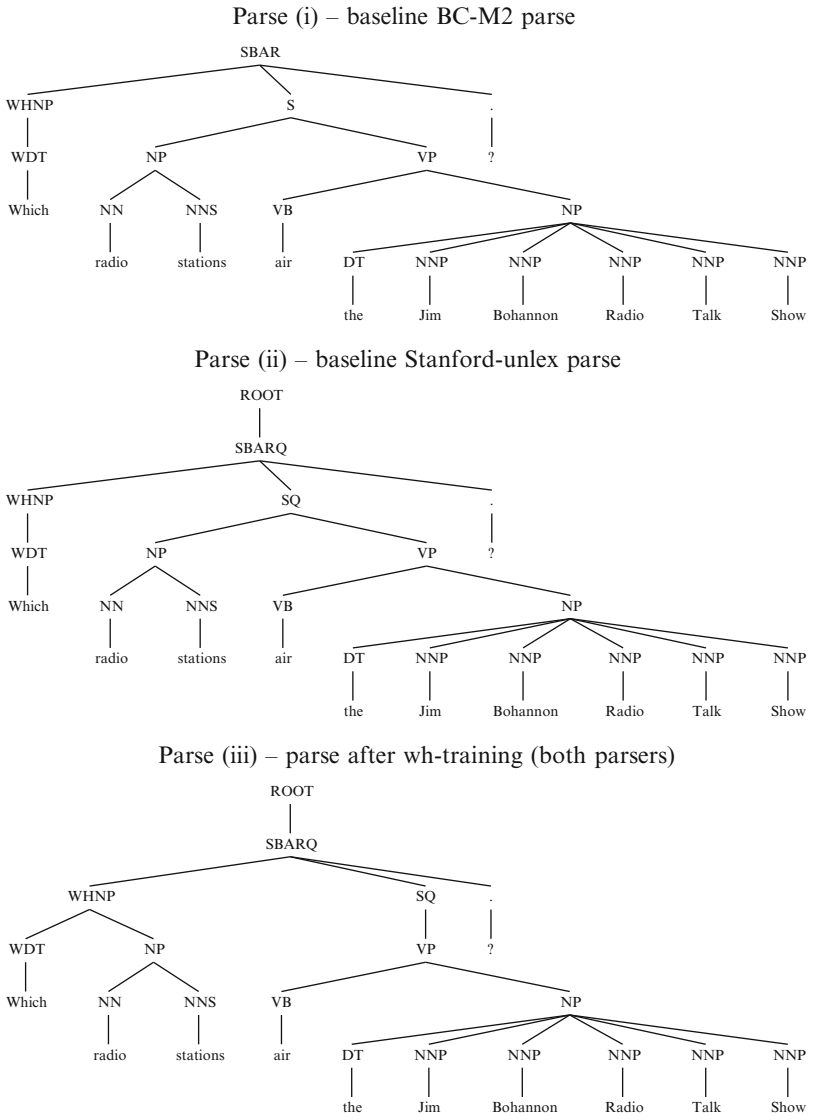

Fig. 5 An example of wh-parsing improvement after wh-training for the test sentence *Which radio stations air the Jim Bohannon Radio Talk Show?* The topmost portion (i) shows the BC-M2 parse before training, with an erroneous S node at the *top*, and the WHNP and NP as distinct trees. Similarly, the Stanford-unlex parse incorrectly separates the WHNP and the NP, while getting the SQ node correct, *middle* display (ii). The *bottom* portion (iii) exhibits the correct parse output by both the BC-M2 parser and the Stanford-unlex and Stanford-fact parsers after wh-training on QuestionBank

wh-training were quite substantial at 6.69 % and 12.31 %, with an overall F-score increase of 9.67 %. Importantly, as the last three columns of the table show, this improvement did not come at any apparent cost in precision/recall for the standard WSJ section 23. For example, the Stanford-unlex parser after additional wh-training got an F-score 85.55 %, on WSJ section 23, as compared to a baseline F-score of 85.54 %. In most cases, the additional wh-examples improve performance.

A representative example of a parse that is greatly improved by wh-training is depicted in Fig. 5, for the test data sentence, *Which radio stations air the Jim Bohannon Radio Talk Show?* Before wh-training, none of the parsers could correctly analyze this sentence. For instance, as expected, the Bikel-Collins parser mis-analyzes the words *which radio stations* as an S dominated by an SBAR, and also mis-parses *which radio stations* as distinct *WHNP* and *NP* phrases (part (i) of the figure). The Stanford-unlex parser does better, without any wh-training; it parses the sentence correctly as an SBAR dominating an SQ. However, it also fails to combine *which radio station* into a single wh-phrase (see (ii) in the figure). After training, both parsers produce 100 % gold-standard parses, shown at the bottom of Fig. 5, panel (iii).

We conclude that the 3,200 questions in QuestionBank, provide a substantial performance boost to wh-question parsing, enough to overcome any deficiencies in the original PTB. However, we note that this puts to one side the question of linking wh-elements with their "underlying" argument structure, as noted by Rimmell et al. [44], among others. In this sense, the fundamental representational question is still not addressed.

4 Parsing and Tense: The Case of *Read*

In a Linguistic Society of America pamphlet, Ray Jackendoff [24] considered a "text reading" puzzle as an example of what is impossible for a computer to accomplish without knowledge of language: in particular, the task of determining the pronunciation of the orthographic form *read*, which can be pronounced as *red* or *reed* depending on context. The sentences considered by Jackendoff are reproduced in Ex. 4; we will consider additional examples as well. In these examples, [24] introduced *will* as a deliberate complication since it can be either a Noun or Modal verb. Apparently, this was to illustrate that simply looking at adjacent words, without any sophistication, would be problematic. In any case, if this issue arises at all, we dealt with it by substituting *should* or *stock* for *will*, as appropriate. The results remained the same, so for our purposes this additional complication was ignored in what follows.

(4) a. The girls will read the paper. (*reed*)
b. The girls have read the paper. (*red*)
c. Will the girls read the paper? (*reed*)

Table 3 The Penn Treebank verbform tagset

Tag	Description	*Example*
VB	Verb, base form	*write*
VBD	Verb, past form	*wrote*
VBG	Verb, gerund or present participle	*writing*
VBN	Verb, past participle	*written*
VBP	Verb, non-3rd person singular present	*write*
VBZ	Verb, 3rd person singular present	*writes*

d. Have any men of good will read the paper? (*red*)
e. Have the executors of the will read the paper? (*red*)
f. Have the girls who will be on vacation next week read the paper yet? (*red*)
g. Please have the girls read the paper. (*reed*)
h. Have the girls read the paper? (*red*)

It should be clear from the examples in (4) that a computer program needs to possess knowledge of the English auxiliary/main verb system along with basic properties of sentence phrase structure in order to correctly carry out this task. The PTB assumes a part of speech tagset that identifies and distinguishes among different forms of a verb, as shown in Table 3. This information ought to suffice, since these values are enough to fix a deterministic decision procedure to pronounce *read* correctly. Note that such a parsing system must be able to associate, e.g., the tense marking on a word like *will* with the correct tense of the verb *read* that appears later in the sentence. General agreement phenomena such as this have been a staple of linguistic analysis for more than 60 years [8]. A related issue appears with other verb forms such as *cut* or *cost*, that are ambiguous with respect to their tense information in the third person (e.g., *they cut*/*they have cut*). In this case, though their pronunciation is also identical, there is still a problem in picking the right tense label for the verb, as we shall see.

One might reasonably expect a parser trained on nearly 40,000 sentences to have acquired basic English sentence structure and properties of the auxiliary and verbal system, and thus be able to decode the examples correctly identifying the appropriate tag for *read* in each case, thus solving the "text reading machine problem" posed by Jackendoff. This is the question we shall examine here.

For example, the structure recovered by the Berkeley parser in the case of 4(b), correctly identifying *read* as VBN, is given in Fig. 6 on the left. (In the case of *read*, only the VBD and VBN forms should be pronounced as *red*.)

However, the Berkeley parser is not always correct. The bottom part of Fig. 6 illustrates the corresponding Berkeley parse for 4(h). Here the sentence has been properly identified as an interrogative (category label SQ) but the parser nonetheless has fails to assign the correct VBN tag to *read*. (The assigned tag VB will result in a pronunciation of *reed*.)

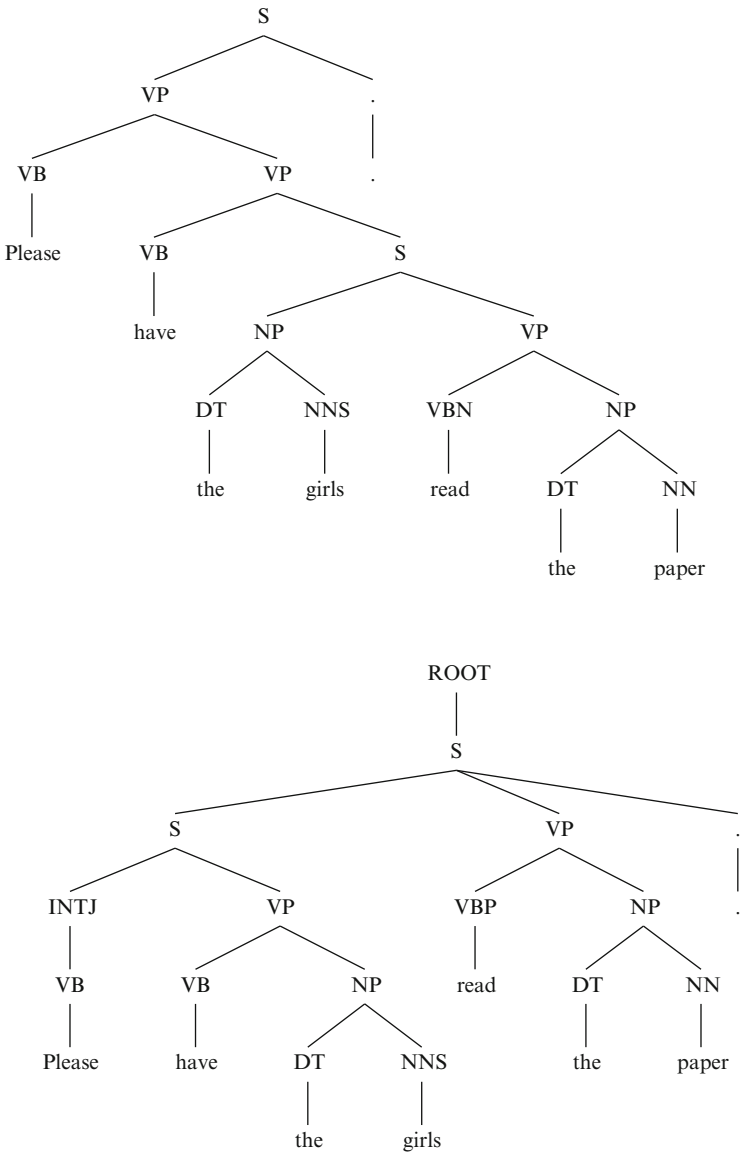

Fig. 6 Berkeley (*top*) and BC-M2 (*bottom*) parses for sentence Examples 4(b,h)

Continuing with this experiment, we examined in some detail how the Jackendoff *read* sentences are analyzed by our suite of statistically-based parsers, all trained on the same sections of the PTB. The results are summarized in Table 4. There are striking differences in performance. Even some of the output parse structures are different. (See Fig. 7 below for a display of a parsing difference with the imperative

Table 4 Parsing results for the *read* pronunciation task. All parsers trained on identical data. Incorrect outputs are flagged with an asterisk*

Example	(4a)	(4b)	(4c)	(4d)	(4e)	(4f)	(4g)	(4h)	
Correct form	VB	VBN	VB	VBN	VBN	VBN	VB	VBN	Correct
Berkeley	VB	VBN	VB	*VB	*VB	*VB	VB	*VB	4/8
BC-M2	VB	VBN	VB	*VB	*VB	VBN	*VBN	* VB	4/8
CJ-I	VB	VBN	VB	*VB	*VB	VBN	*VBN	*VB	4/8
CJ-R	VB	VBN	VB	*VB	*VB	VBN	*VBN	*VB	4/8
Stanford-unlex	VB	VBN	VB	VBN	VBN	*VB	VBP	VBN	7/8
Stanford-fact	VB	VBN	VB	VBN[a]	VBN	VBN	VBP	*VB	7/8

[a]This assumes that the parser has not misinterpreted *will* as a modal verb. The same holds for the next example

sentence Ex. 4(g).) Overall, the Berkeley parser gets 4/8 of the test sentences correct, missing 4(d–f,h).[6]

The BC-M2 parser does not have perfect performance either, with 4/8 correct, though it fails on a slightly different set of examples; it misses 4(d,e,g,h). For comparison, note that an assignment based purely on tag frequency would yield a crude baseline of 3 out of 8 correct on this task, as VB and VBN occur 45 % and 19 % of the time in the training set for *read*. It is important to observe that unlike the other parsers tested here, the BC-M2 parser ignores final sentence punctuation, so it literally cannot distinguish *Have the . . .?* from *Have the . . .*.

The other two lexicalized parsers, both the 'first-stage' *n*-best parser using Charniak's "coarse to fine" method and the CJ re-ranking parser, perform exactly the same as BC-M2, getting 4/8 sentences right, and missing the same sentences as BC-M2, on sentences 4(d,e,g,h).[7]

Finally, turning to the two Stanford parsers, we see greatly improved performance. If we count VBP as OK for the imperative *read* sentence, then the (simpler)

[6]As noted in Sect. 2 we tested both the Berkeley's parser's pre-built `eng_sm5` grammar, as well as our own retrained version that carried out six split-merge iterations. The results did not change. The results also remained the same when we used Berkeley parser's `-accurate` switch. In general, results did not change for any of the parsers when we substituted *stock* or *should* for *will*. Note that here the Berkeley parser is using its own part of speech tagger. If we force it to use "gold standard" part of speech tags, then it could not possibly fail in the manner we have described. However, we wanted to examine the parser's own performance, not some exogenous part of speech tagger.

[7]For CJ-I we selected the "best" (highest likelihood parse score) from the output of the CJ-I parser. In fact, in several cases, the 2nd best parse tree turned out to be the correct one; this was true, for instance, for sentence 4(h). On the other hand, just as often the best parse was correct and the 2nd best parse was incorrect, as in example 4(a). Note that the CJ-I parser serves as input to the CJ-R re-ranking parser, taking, e.g., the top-50 most likely parses and then sorting them according to a discriminative weighted feature-based scheme using features such as the degree of right-branching, or conjunct parallelism. Since the top 50 parses usually included the correct answer, the re-ranking parser at least had a chance of possibly selecting the correct answer in each case. Even so, re-ranking was ineffective, and did not change the outcome for any of the sentence examples here. See [6] for details about this re-ranking parser.

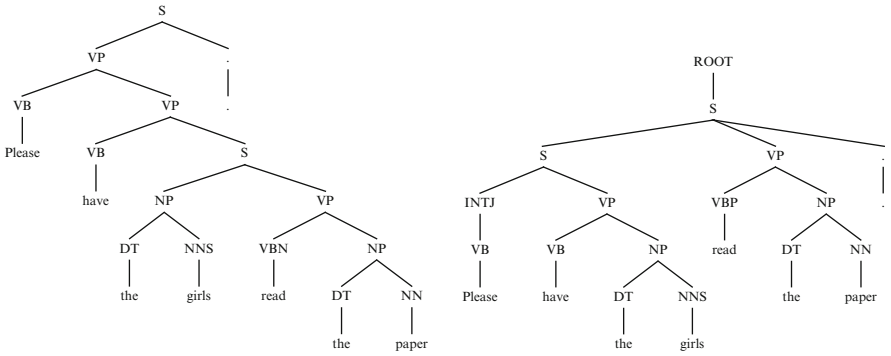

Fig. 7 Some parsers output distinct structures for the imperative *read* sentence. The *left-hand* side displays (identical) the parse output by the Berkeley, and BC-M2 parsers. (The CJ-I and CJ-R parses are also identical to this one, aside from the minor difference of labeling *have* as an AUX.) The *right-hand* side displays the output from the Stanford parsers for this same sentence

Stanford unlexicalized, probabilistic context-free parser is nearly perfect, with 7/8 sentences correct. The more sophisticated dependency-factored Stanford parser also gets 7/8 correct, (Both of these parsers also output different, arguably incorrect parses for *Please have the girls read the paper*, displaying the imperative form as shown on the right-hand side in Fig. 7.)

What accounts for the difference in the results? All of the parsers use extremely sophisticated statistical estimates, with many programming details, so it is very challenging to determine what accounts for their varying performance on particular sentences. As Bikel observes, [4], p. 188:

> With so many parameters, a lexicalized statistical parsing model seems like an intractable behemoth. However, as statisticians have long known, an excellent angle of attack for a mass of unruly data is exploratory data analysis.

We shall pursue such an exploratory path here. Let us consider first the essentially identical performance of the BC-2, CJ-I, and CJ-R parsers. As noted in [5], all these parsers are strongly "lexicalized," in the sense that they use literal word information about the heads of phrases in the linguistic sense (smoothing this if necessary by various methods). That is, instead of a rule expanding a Verb Phrase (VP) as $VP \rightarrow VNP$, these parsers modify the context-free rule to incorporate actual information about the lexical head word, e.g., the particular verb *read*. The by-now familiar advantage here is to possibly capture any special properties that distinguish *read*, from, say, *buy* – perhaps that *buy* is more frequently followed by an object Noun Phrase. Such systems thus serve as a point of contrast with the remaining parsers tested, which do not in general expand context-free rules with augmented head information. We put to one side for now the method that the factored Stanford parser uses, which is in effect to parse with both an ordinary PCFG and a lexicalized dependency model, and then combine the results by means of a joint inference model.

More specifically, we may be able to pinpoint the difficulty with the lexicalized parsers by drawing on an observation made by Charniak [5]. Charniak notes that the BC-M2 parser and the CJ-I and CJ-R parsers all make use of actual lexical information, to first "guess" whether a pre-terminal label should be, e.g., VB or VBN, p. 137:

> . . . the current parser first guesses the head's pre-terminal, then the head, and then the expansion. It turns out that usefulness of this process had already been discovered by Collins [14]. . . However, Collins . . . does not stress the decision to guess the head's pre-terminal first, and it might be lost on the casual reader. Indeed, it was lost on the present author until he went back after the fact and found it there.

While [5] notes that this method accounts for a nearly 2 % performance gain overall, there is some evidence that it also leads to precisely the observed problem with *read*, essentially one of "over-lexicalization." In particular, as explained in [3], the BC-M2 parser "guesses" the part of speech of a pre-terminal associated with *read* via a top-down generative approach, sometimes modifying the pre-terminal part of speech information. We can see the effect of this in the case of *read*. In the example *Have the executors of the will read the paper*, *read* is initially assigned the (correct) part of speech tag VBN by a pre-processor tagging step. But this is changed by the probability model's guess of the incorrect tag VB. Indeed, the same holds for the other mistakes BC-M2 makes: initially correct tags are changed to their incorrect counterparts by the parser.

Our hypothesis, then, is that the local "guessing" carried out by the generative probability model in these cases may be biased by local frequency effects in such a way as to sometimes alter the tag in the wrong direction. For example, *read* appears in the PTB training data 29 times as a VP dominating a VB (usually with an intervening *to*), and 10 as a VBP, so in 39 contexts is pronounced *reed*. On the other head, *read* appears 24 times dominated by VBD or VBN, pronounced *red*. It is this bias that appears to be altering the results. In contrast, consider the tense-ambiguous verb *hit*, which appears 88 times as VBD/VBN and only 23 times as a VB/VBP. This distribution is the converse of *read*. Running the same sentences as in 4 through the parsers with *hit*, instead of *read*, e.g., *Have the girls who will be on vacation next week hit the paper*, we find that the number of mistakes is reduced, with the correct tag VBN replacing the incorrect VB tag in three cases. Similarly, *cost*, which has the same rough local frequency distribution as *read*, with 65 VB/VBP and 22 VBD/VBN counts, behaves as expected like *read*; so does *cut*. If this view is on the right track, then it is these local frequencies, which are sensitive to the small sampling effects of the PTB, that are at play here. Further, this same issue seems to infect the other two "lexicalized" parsers, though not to precisely the same extent: when we replace *read* with *hit*, then the CJ-I and CJ-R parsers now get sentences 4(d,e) correct (as does BC-M2), but these two parsers still fail on the last two sentences. Some kind of lexicalization effect is operating, but it is not exactly the same as that with BC-M2, perhaps because the CJ parsers augment the standard PTB part of speech categories with the addition of AUX for *have*.

Additional confirmation of the effect of lexicalization comes from examining the behavior of the unlexicalized parser, Stanford-unlex. It does not make any assumptions about lexical heads, and so we would not expect it to be subject to the variation we see with the lexicalized parsers. In fact, as shown in Table 4, it is much more successful, making only one mistake, labeling *read* as a VB in *Have the girls who will be on vacation next week read the paper yet*. Note that the addition of a lexicalized component that is grounded on dependencies, the factored Stanford model that uses both word dependencies and the Stanford unlexicalized parser to jointly infer structure, also makes a single error, but it is not the same one. Instead, it makes an error on the last *read* sentence, taking it as a VB rather than a past-tense VBD. While the reasons for these singleton errors remain obscure, it is clear that this approach works better than straight lexicalization.

It remains to account for the behavior of the Berkeley parser. While it is not lexicalized, it works by refining categories and rules by successive state-splitting. It may be that its "window size" for learning context is too narrow. The trainer uses a context window based on horizontal (h) and vertical (v) "markovization," that is, how many past horizontal ancestors are remembered, and how many vertical (parent, grandparent) ancestors are remembered, as a context for future parsing decisions. By default, these values are set to 0 and 1, respectively – that is, a context that remembers only the immediate parent node above a current position. Note that in an imperative form like 4(g), the "distance" between the verb *have* and *read* lies outside this window. In [27], larger values for h and v are systematically explored, with some evidence provided that h and v values larger than 0 or 1 may be needed for generally effective performance. It remains to explicitly test this hypothesis precisely within the context of the *read* example.

How can we improve the performance of the parsers on the *read* examples? If the effect is due to sparsity and lexicalization, then as with the wh-question case, more data might prove helpful. Here the models distributed with the Stanford parser themselves indicate that additional data of the right kind indeed can be a benefit. Along with models trained solely on the PTB, Stanford-unlex and Stanford-fact come with models trained on a selection of biological abstracts from the GENIA corpus [51], plus 96 "additional" hand-built parse trees; these are called `englishPCFG` and `englishFactored`. Importantly, the 96 "additional" hand-labeled examples include examples that are directly comparable with the *read* examples, including 11 relatively short subject questions, SQs typically with subject-auxiliary verb inversion, such as *Is what she said untrue*; and 25 wh-questions, or SBARQs, such as *Where was the fox*.[8]

Probing a bit further, if we run the *read* examples using the Stanford models based on this augmented corpus then they do perfectly, so it would seem worthwhile

[8]The remaining examples are some simple S's and a few newswire stories. The authors would like to thank C. Manning for generously sharing these additional examples with us.

Table 5 Parsing results for the *read* pronunciation task when rerun on non-Stanford models retrained on the augmented PTB + Stanford "additional examples." Errors are marked with asterisks, as before

Example	(4a)	(4b)	(4c)	(4d)	(4e)	(4f)	(4g)	(4h)	Correct
Berkeley	VB	VBN	VB	*VB	VBN	VB	VB	*VB	6/8
BC-M2	VB	VBN	VB	*VB	VBN	VB	VB	*VB	6/8
CJ-I	VB	VBN	VB	*VB	*VB	VBN	VB	*VB	5/8
Stanford-unlex	VB	VBN	VB	VBN	VBN	VBN	VBP	VBN	8/8
Stanford-lex	VB	VBN	VB	VBN	VBN	VBN	VBP	VBN	8/8

to examine what is causing the improvement, as was true in the wh-question case study. To examine this, we tested whether the 96 extra examples alone would suffice to correct some or most of the *read* errors. We therefore retrained all the parsing models, aside from CJ-R, using just the PTB training data plus the 96 "additional" examples, omitting the GENIA examples. We then re-ran the *read* example sentences, with the results shown in Table 5. There is an improvement in every case. Both Stanford parsers still have perfect scores, suggesting that the entire improvement is due to the 96 extra examples, rather than further additions from GENIA. Further, both the Berkeley, BC-M2. and CJ-I parsers improve, and now get 6/8 correct (they all fail on the third and the last *read* examples). We conclude that the judicious addition of even a few critical examples can greatly improve parsing performance, just as in the case of QuestionBank, again pointing to the sparsity of the original PTB training dataset as well as the ease with which some its failings may be remedied, at least in this particular situation.

However, it is still true that none of the systems explored here explicitly records the linguistic fact that the auxiliary at the front of the sentence is tied to the main verb. They do so only indirectly. Even in English, the properties of tense are "spread out" over the entire Auxiliary system. In an example such as *The stock could have been being sold*, it is the sequence of auxiliary verbs that together carry the tense information. It is only a morphological accident of English that these elements must generally be string-adjacent. Whenever two are separated by an intervening phrase, as in the *read* examples, the agreement between them still holds. It remains to be seen how to properly represent such facts in the statistically-grounded systems we have explored here.

Here we note that parameter estimation issues are a symptom rather than the underlying cause of the deficiencies of the parsing model. Such a model is unable to capture the interaction between wh-movement and the auxiliary/main verb system, or posit a connection from the declarative form of the sentence to its interrogative form without actually having observed the handpicked examples that closely match the test data.

5 Case Study: Parsing Passives by Linguistic Regularization

We noted in Sect. 1 that statistically-trained parsers make attachment errors in passive sentences, in part because attachment decisions are difficult without sufficient data. We also pointed out that in certain cases, this could be repaired by reconstructing a sentence's underlying "logical form" (a form of "D-Structure" in the classical sense), thereby rendering arguments in canonical positions. In general, we will call these kinds of reconstructions into a canonical predicate-argument form *linguistic regularizations*.

We note that several researchers have previously attempted to improve statistical parsing performance via representational changes to the grammar, in the form of either tree-level transformations, or by incorporating other latent information present in the Penn Treebank [7, 19, 25, 32]. Most of these approaches follow the paradigm proposed in [25], whereby the parser is retrained on a transformed version of the training set and then after evaluation the resulting parses are de-transformed and evaluated against the known gold standard annotations.

The approach we will take here differs from this past research in at least two critical respects. First, previous work such as that in [30] has focused on using additional features in the PTB as a means to improve parsing accuracy, while still others, as in [15] Chap. 7, model wh-displacements by means of feature passing. Few approaches have explicitly modeled a separate level of underlying predicate-argument structure. Second, more specifically, the level of syntactic complexity involved in these transformations has been rather limited, and none of the researchers up to the present point have attempted to reassemble the underlying representation of passive constructions.

Following the methodology of [25], we propose to exploit the additional information provided by linguistic regularizations in the following way. First, as suggested above, we can use the annotated PTB training trees to "invert" various displacement operations, returning arguments to their canonical "underlying" positions. In the case of our example sentence, we would derive something like, *Tablespoons may soon replace measuring cups in the laundry room*. We then use the transformed sentences as revised training data for a statistical parser. If the regularization idea is sound, then we would expect improved performance.

5.1 Passive Transformations: A Pilot Study

We will now show that employing "logical form" structural cues for linguistic regularization can improve parsing performance within the existing Penn Treebank formalism. We selected the passive because it has not, to our knowledge, been tackled in previous work. The experimental setup is as follows. As mentioned, we approach the problem within the framework proposed by Johnson [25]. We identify a set of transformations we would like to model in the corpus, transform the input

Table 6 Parsing results for models trained on the original (BASE) and transformed (TRANS) Penn Treebank (PTB) data. *untrans* corresponds to the untransformed or original corpus, while *trans* to the transformed version. *full* is the entire corpus; *psv*, the subset of passive sentences; *yactive*, the subset of active sentences. SBASE and STRANS experiments are oracle experiments – where the test set ("special") sentences are selectively transformed or kept intact to maximize the evalb recall. The POS column corresponds to the part of speech tagging accuracy. The size column identifies the number of sentences in the test corpus

Experiment id	Training set	Test set	Recall	Precision	POS	Size
BASE-1	wsj-02-21 untrans	wsj-23-full-untrans	88.17	88.36	96.87	2,416
BASE-2	wsj-02-21 untrans	wsj-23-full-trans	87.89	88.08	96.73	2,416
BASE-3	wsj-02-21 untrans	wsj-23-psv-untrans	87.75	87.96	97.40	364
BASE-4	wsj-02-21 untrans	wsj-23-psv-trans	86.28	86.43	96.65	364
BASE-5	wsj-02-21 untrans	wsj-23-active	88.27	88.45	96.75	2,052
TRANS-1	wsj-02-21 trans	wsj-23-full-untrans	88.26	88.48	96.86	2,416
TRANS-2	wsj-02-21 trans	wsj-23-full-trans	88.29	88.47	96.82	2,416
TRANS-3	wsj-02-21 trans	wsj-23-psv-untrans	87.39	87.65	97.27	364
TRANS-4	wsj-02-21 trans	wsj-23-psv-trans	87.51	87.62	97.02	364
TRANS-5	wsj-02-21 trans	wsj-23-active	88.46	88.66	96.77	2,052
SBASE	wsj-02-21 untrans	wsj-23-psv-special	88.12	88.22	97.02	364
STRANS	wsj-02-21 trans	wsj-23-psv-special	89.30	89.38	97.25	364

data by performing a set of deterministic 'tree' surgeries on the input parse trees, and then, after re-training, evaluate the resulting parser on a transformed test set.

The first step in this process is to perform tree regular expression (`tregex`) queries on the corpus to identify the passive constructions in the training data sections of the PTB. Second, we must map passive syntactic structures back into their active form counterparts. This mapping is achieved through a sequence of tree-transforms, applied recursively in a bottom-up, right to left fashion using the `Tregex` and `Tsurgeon` toolkit [31]. Note that in some cases, there will be no "by" phrase, that is, no explicit semantic Subject. In these cases, we insert a dummy subject with the part of speech label `TT`, corresponding roughly to *it*.

In all, there are 6,015 passive sentences in the training corpus out of a total of 39,832 sentences, or 15 % of the training data. In the test set, section 23 of the PTB corpus, 364 out of 2,416 sentences or 15.1 % of the test data can be identified as passives, comparable to the figures observed in the training set. The passive construction would therefore seem to provide a good test-bed for a pilot analysis. A ten percent sample of the identified training set items and all of the test set items were manually checked by a human expert who validated them as true passive constructions.

The third step of the procedure is to re-train and test a statistical parser on the transformed test and training data. We conducted our experiments using BC-M2 [3], following standard procedures. Additionally, we conducted our experiments on different combinations of transformed and untransformed training and test data, as well as allowing for configurations whereby the test corpora were evaluated on the active and the passive subsets separately. The pilot test results are displayed in Table 6.

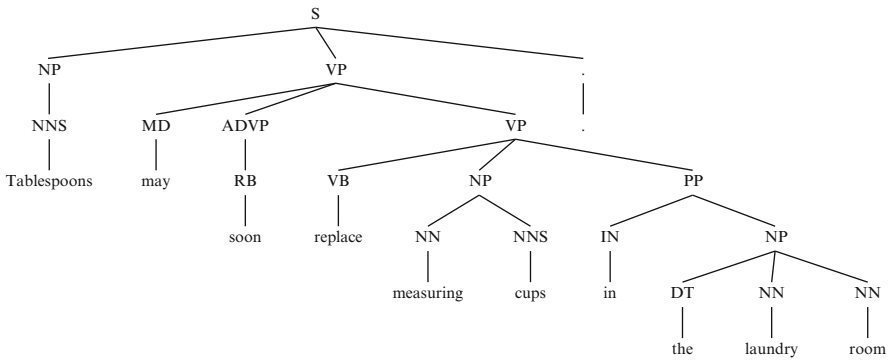

Fig. 8 The Bikel/Collins parser correctly analyzes the "tablespoon" sentence after regularization

First, we note that the baseline parser (BASE-*) performed markedly better on the active sentence set than on the passive construction subset of the WSJ corpus section 23 (88.27 % vs. 87.75 % recall). This lower score is to be expected, since the passive construction exhibits longer-range movement and constitutes only 15 % of the training data.

On the full test set (2,416 trees), the retrained model (TRANS-2) beat the baseline (BASE-1) by 0.12 % absolute recall (88.29 % vs. 88.17 %) and 0.11 % absolute precision. On the active sentence subset that constitutes about 85 % of the test corpus, the model outperforms the baseline by 0.19 percent in recall – a statistically significant difference at the 0.05 level (p-value = 0.029) as computed by a stratified shuffling test with 10,000 iterations. While this may seem like a small performance gain, in the context of a trained parsing system that is known to be operating at close to a theoretical ceiling, this is in fact a real performance increase.

More concretely, to give an idea of an error that is corrected by regularization, in Fig. 8 we display the parser's output of the transformed example sentence, *Tablespoons may soon replace*... The parser outputs a tree that is 100 % correct.

To give a broader picture of where the performance improvement comes from, as another example, Fig. 9 displays an example from section 23 of the PTB, sentence #722, *According to analysts , profits were also helped by successful cost-cutting measures at Newsweek* ., that is parsed incorrectly in its unregularized form, with a misplaced PP high attachment for *at Newsweek*. This yields a labeled precision score of 91.67 % and a labeled recall score of 84.6 %. As the bottom half of Fig. 9 shows, after regularization this sentence is now parsed with perfect recall and precision, with a correct PP attachment under the NP.

Many other mis-parsed passives from the test dataset are parsed correctly after regularization. In all, out of 364 test sentence passives, 74 improved after regularization. Many of these improvements appear to be due to correction of mis-analyzed PP attachments, as anticipated.

However, the simple regularization carried out in the pilot study can sometimes also lead to worse performance: 95 out of 364 test sentence passives were

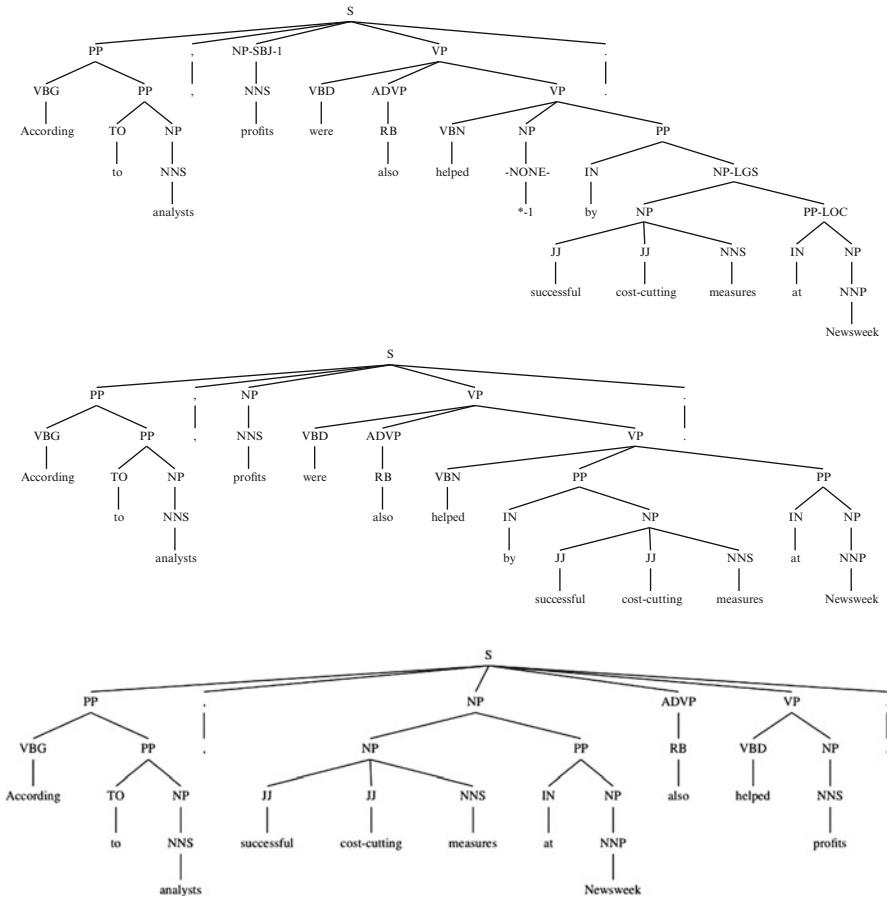

Fig. 9 The BC-M2 parser mis-analyzes of sentence #722 in section 23 of the PTB. The *top* third of the figure shows the gold standard parse. The *middle* third of the figure displays the corresponding (incorrect) BC-M2 parse. The *bottom* third shows the result of parsing the same sentence correctly after the regularization procedure described in the main text

parsed *worse* than before. It is these cases that reduce the performance gain of regularization in our pilot study. Figures 10 and 11 illustrate one example of this effect. Sentence #2,274 in test section 23, the passive sentence, *Tandem 's new high-end computer is called Cyclone*, is parsed with perfect precision and recall before regularization, though with an arguably incorrect gold-standard bracketing: both an empty Subject NP followed by a predicate NP *Cyclone* are dominated by an S. As Fig. 11 shows, after regularization, the re-trained parser mis-analyzes this structure with both the restored Subject NP *Tandem 's* and the predicate NP *Cyclone* combined as a single NP (precision = 71.43 %, recall = 83.33 %). It seems likely that examples such as these might be successfully analyzed if the gold-standard was assigned a linguistically more accurate "small clause" type structure.

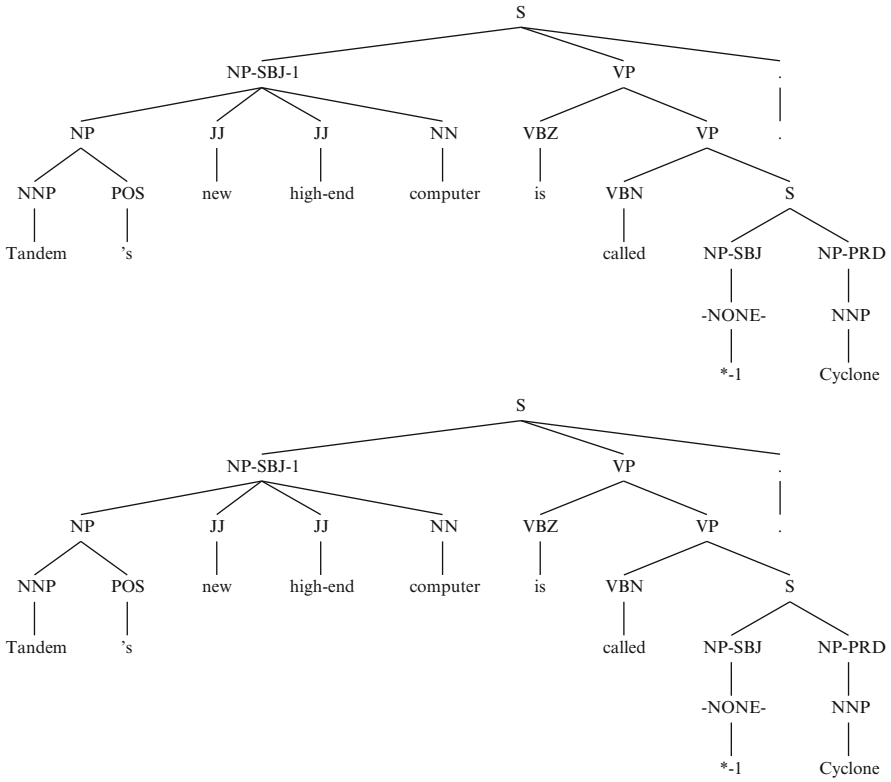

Fig. 10 The Bikel/Collins parser analysis of sentence #2,274 of section 23 of the PTB. The gold standard annotation is at the *top*, the parser output on the *bottom*

Other regularization failures occur where there is no following PP phrase in the original sentence to be mis-parsed, and where the regularization leads to a complex structure with the potential for misanalysis. For instance, the section 23 passive sentence #269, *The land to be purchased by the joint venture has n't yet received zoning and other approvals required for development , and part of Kaufman & Broad 's job will be to obtain such approvals* . requires the NP *the joint venture* to be restored as the Subject of *receive*. However, the re-trained parser incorrectly analyzes the regularized sentence. In part this may be the result of not completely reconstructing the underlying form; in this instance, where there is a relative clause *the land purchased by the joint venture*, the object of *receive*, *the land*, is not explicitly restored to its underlying position after the verb. Such complexity has tendency to lead to mis-analysis, and a more complete reconstruction of such relative clauses might repair such instances.

Note that even though on the passive subset (364 trees) the baseline outperforms the transformed model by 0.24 % recall, the result is not statistically significant (p-value = 0.295). Taken together, the results indicate that retraining significantly

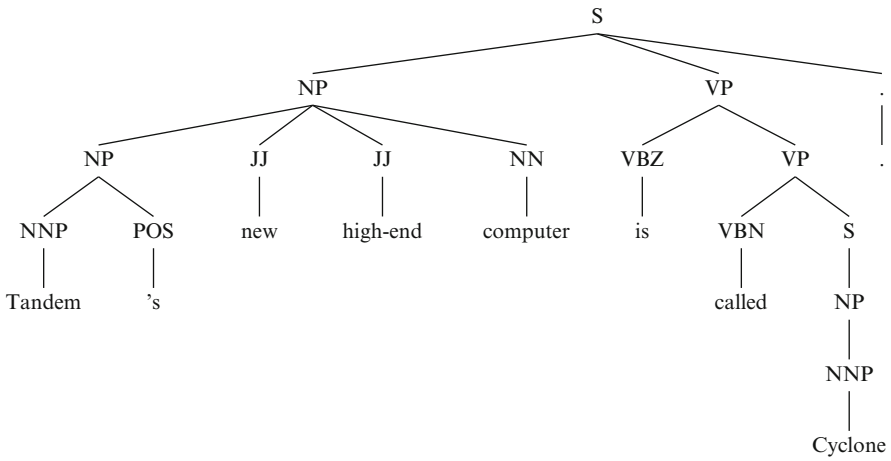

Fig. 11 The parse of regularized sentence #2,274 mis-analyzes the NP – NP structure under a single NP, precision = 71.43 %, recall = 83.33 %

improves the performance of the parser on active sentence constructions, while not incurring a statistically significant loss on passives. In fact, the retrained model is much more robust with respect to untransformed passives, only exhibiting a 0.12 % loss in precision, whereas the baseline suffers almost a 1.5 % degradation (TRANS-3 vs. TRANS-4).

We tested further potential for improvement by selectively unwinding certain passives into their underlying logical form, while leaving others in their original surface form. This is an oracle experiment, whereby we evaluate the parser only on the surface forms that achieve better performance under the retrained parsing model. That is, we assume the presence of an "ominiscient" selection procedure that allows us to decide whether the instance to be parsed for testing first needs to be transformed or whether it is more desirable to leave it in its original form. In carrying out the experiment we evaluated both forms for each test sentence and picked the one that achieved maximum evalb recall. Note that in practice, we would not have access to such a procedure. However, it is instructive to carry out this experiment, as it allows us to gauge the best possible (upper bound) performance for using an "unwound" logical form. This result indicates that we can obtain an upper bound of 89.30 % recall, as much as a full percentage point improvement over the baseline by applying the transformations on a selective basis. Further analysis of the results shows that this effect is achieved due to cases where displaced modifiers in the passive construction impact negatively on the parser's attachment decisions.

Based on the evidence from the oracle experiment, we hypothesize that a simple binary classifier that could choose the training model from the features of the input test sentence should be able to recover much of the hypothetical gain due to the oracle.

Although seemingly small, the improvements obtained in the regularization experiments are statistically significant, and with more engineering effort in modeling nested passives and long-distance displacements we expect a greater gain.

We note that the important takeaway message from this pilot experiment is not that this is exclusively a parameter estimation problem. On the contrary, we point to the impracticality of adding a passive or active instance for every surface form observed in the training corpus without the extra linguistic knowledge explicitly encoded through structural transformations that map passive forms to their active counterparts. By incorporating linguistic knowledge we were able to improve a broken model indirectly by alleviating the parameter estimation problem.

By no means should this fix be viewed as a permanent solution. Our ability to make an impact suggests that the underlying representation is deficient and that much more radical changes need to be made to the model. One approach, by no means the only one, is by explicitly representing movement as a primitive operation. Alternatively, one could adopt a scheme like that of Combinatorial Categorial Grammar.

6 Parsing "Unnatural" Languages?

We turn in our final section to the Musso et al. experiment [36], in an attempt to probe to what extent statistically-based parsers can acquire "unnatural" language constructions. Recall from Sect. 1 that the second experiment in [36] was designed to see whether normal adults could easily learn a "mirror reversed" question formation rule, as well as whether this learning (as tested by subsequent parsing probes) activated the same brain regions, as visualized by fMRI. A typical example of such an natural/mirror-reversed pair, as cited earlier, is this: *il bambini amano il gelato*/*gelato il amano bambini il*. Their basic finding was that normal adults had extreme difficulty with such examples, solving them, if at all, as if they were non-linguistic puzzles, and drawing on different brain regions than those usually seen associated with language (specifically, outside Broca's area). Similar poor learning of "unnatural" language patterns has also been found in autistic language savants [49].

Our last experimental manipulation investigated whether we could replicate the second study described in [36] within the context of statistically-trained parsing. That is, we modified the PTB training data so that all question forms would be presented in their reverse or "mirror image" order, rather than in normal English word order. The parsers would then be trained on this manipulated data, and subsequently tested whether they had acquired the "mirror reverse question" construction by assessing them on a similarly question-reversed PTB section 23 data set.[9] In our emulation experiment, in addition to the standard PTB training

[9] We put to one side the question of carrying out fMRI experiments on computers.

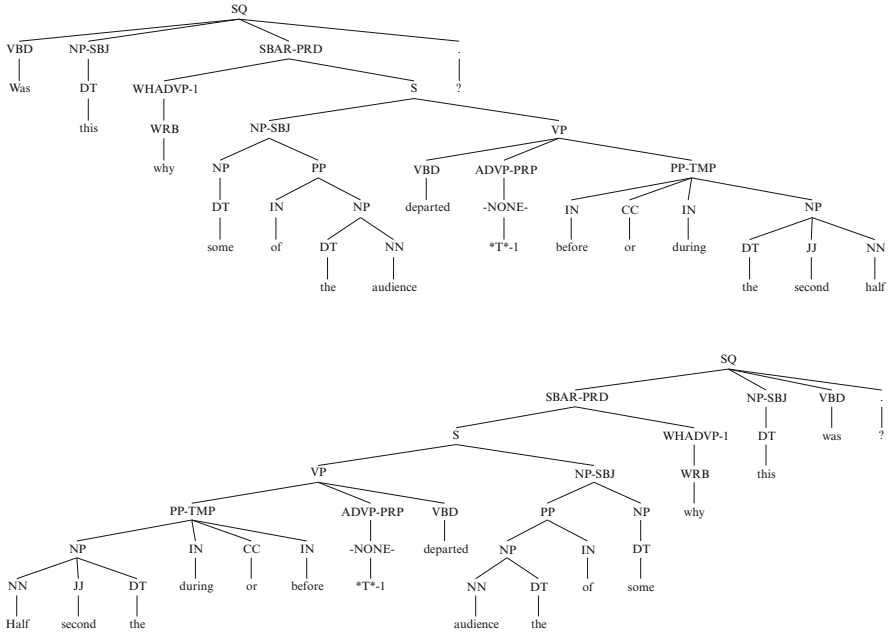

Fig. 12 Conventional and mirror-image treebank questions from the PTB, for training sentence#76, Was this why some of the audience departed before or during the second half?

sections, we also carried out a supplementary training/test regime again using the QuestionBank constructed by Judge et al. [26]. We did this because there are only 24 questions total in the entire standard test section 23 of the PTB, so that mirror-reverse questions are not properly exercised by the normal test dataset.

A typical example of such a "mirror image" training tree drawn from QuestionBank is displayed in Fig. 12 below, the mirror image corresponding to the question, *Was this why some of the audience departed before or during the second half?* Note that the input words are in reverse order (and the parse tree is the mirror reflection of the given parse tree in the treebank).

We should emphasize that there is a considerable challenge in carrying out this exercise properly in order to reflect (as it were) adult linguistic behavior and inference. It is, in general, not possible to exactly replicate the experimental conditions in [36]. The key problem is that we cannot be certain as to the internal system by which people processed the reversed sentences in [36]. As a first approximation, however, it may be fair to say that they could bring to bear the usual cognitive apparatus of "chunking" words into phrases (though the exact manner and details as to how much structural information is readily available remains a matter of some controversy; see [45], among much other recent work on this topic). However, it is reasonable to surmise that they did not have access to pre-formed parse structures, as is the case with the artificially constructed corpuses and the

statistically-trained systems. In particular, in our emulation we gave the parsers the mirror-images of question sentences (including those embedded in quotational contexts), and one might reasonably object that this is far more information than that provided to the human subjects. This is a fair point. However, here we shall simply observe that [36] deliberately used Japanese (and German) native speakers for their experiments, just for this reason, since these languages are head-final, with left-branching structure similar to that displayed on the bottom half of Fig. 12, though of course not so uniformly reversed and not reversed solely with respect to questions. This was intended to compensate for any basic unfamiliarity with branching structures of the kind displayed in the figure, the implication being that these speakers would have had experience grouping lexical items in such a fashion. Further, this is evidence that intonational breaks to highlight structure and related cues are essential in some way for language inference in any case; see [35]. However, there is no denying that the exact experimental condition we used, providing both the reversed string and its corresponding mirror-image parse tree, has, to the best of our knowledge, never been replicated in any human subject experiment. This is true of many important questions regarding human language acquisition. For example, until it was first probed in [17], whether or not children actually formed Subject-Auxiliary verb questions using structural rules had not been experimentally addressed. Similarly, the question posed here is an empirical one that can only be resolved by future research.

6.1 The Experimental Emulation

To emulate the experiment in [36], we prepared two sets of training and test data, all with reversed questions, via manipulation of the PTB, along with the additional QuestionBank corpus. To start then, we had two training and two test datasets: (1) the standard training sections 02–21 of the PTB; (2) test section 23 of the PTB; (3) the normal training sections of the PTB concatenated with an 80 % sample of QuestionBank, 3,200 questions; (4) a held-out 20 % test sample of QuestionBank, 800 questions. (See Sect. 4 for a detailed description of QuestionBank.)

To obtain the appropriate mirror-image "reversed" question datasets we replaced all questions (both root level questions and questions in sentence contexts, usually quotational) in the original corpuses with their mirror-image counterparts. Figure 12 displays an example of a PTB training sentence #76 in its normal and mirror-reversed formats. The original sentence is, *Was this why some of the audience departed before or during the second half?*, while the reversed structure corresponds to, *Half second the during or before departed audience the of some why this was?* An example of a wh-question in a quotational context is sentence #610 of the training set, *"So what if you miss 50 tanks somewhere?" asks Rep. Norman Dicks, D., Wash., a member of the House group that visited the tanks in Vienna.* We carefully analyzed the original data to ensure that these were properly reversed. In this case, only the material within double quotes would be reversed.

For convenience, we will refer to all these training and test data sets along with their mirror-image question reversed counterparts as follows. There are four training sets in all, the two non-question reversed training sets and the two question reversed training sets. Similarly, there are four corresponding test sets. So altogether there are a total of 16 possible training-test dataset combinations. We will denote each of these training/test combinations with a unique label consisting of the training dataset name, a slash, and then the test dataset name. For example, WSJ/WSJT denotes the conventional WSJ training/WSJ section 23 test combination, while WSJR-QBR/QBRT denotes the WSJ training section with mirror-image questions augmented by the mirror-image questions as the training set, and the held-out mirror-image QuestionBank sentences as the test set. Note that the QuestionBank and the WSJ corpora are disjoint. The four training and four test sets are as follows.

1. **WSJ:** The conventional training sections 02–21 of the PTB;
2. **WSJR:** The question mirror-reversed training sections 02–21 of the PTB
3. **WSJ-QB:** The question-augmented corpus, sections 02–21 + the 80 % sample from QuestionBank;
4. **WSJR-QBR:** The question-reversed WSJ training section + mirror-reversed QuestionBank 80 % sample;
5. **WSJT:** The conventional test section 23 of the PTB;
6. **WSJT-R:** The question-reversed conventional test section 23 of the PTB;
7. **QBT:** The 20 % held-out test sample from QuestionBank;
8. **QBRT:** The question-reversed sentence test sample of QuestionBank.

6.2 Training, Testing and Results

We selected the BC-M2 and Stanford-unlex parsers as representative "lexicalized" and "unlexicalized" parsers for the experiment. Along with 16 training-test combinations, this yields 32 possible experimental runs. Note that four of these runs, the WSJ/QBT and WSJ-QB/QBT analyses for each parser, have already been carried out as part of the wh-QuestionBank testing in Sect. 3, but we include them below for completeness.

The results are summarized as F-scores in Tables 7 and 8. (We have split the results across two tables in order to highlight the most important contrasts in the first table.) The first table's results are also displayed in a more readable form as the histogram in Fig. 13, which presents F-scores on the Y-axis, and the most important training-testing contrasts on the X-axis; the BC-M2 results are in dark grey, and Stanford-unlex in light gray. Note that because there are so few questions in test section 23 of the PTB, just 20 out of 2,416 sentences, excluding a few non-question fragments that are marked as questions, that performance on the WSJ-T corpus does not serve as a reliable indicator of whether question sentences have been learned or not, though it may be of some value to see whether learning mirror-questions

Table 7 F-score results for the first eight training/testing results for the "mirror reversed" experimental manipulation. Lines (4)–(7) show that both lexicalized and unlexicalized parsers learn "mirror reversed" questions quite well

Train-test combination	BC-M2	Stanford-unlex
(1) WSJ/WSJT	85.63	85.54
(2) WSJ/WSJT-R	85.78	85.71
(3) WSJ/QBT	75.76	67.75
(4) WSJ/QBRT	13.15	19.12
(5) WSJR/QBRT	58.04	61.20
(6) WSJR-QBR/QBRT	65.94	71.47
(7) WSJR-QBR/QBT	55.67	60.58
(8) WSJ-QB/QBT	86.18	81.32

Table 8 The remaining 16 results for the WSJ "unnatural" learning experiments. Note that training by reversing just the questions in the WSJ, using WSJR, also boosts reversed-question parsing performance, but not as much as using the full training QBR training set. In general, testing on WSJR does not indicate any great difference, because there are so few questions in WSJT to test

Train-test combination	BC-M2	Stanford-unlex
(1) WSJ-QB/WSJT	85.79	81.32
(2) WSJ-QB/WSJT-R	88.01	85.46[a]
(3) WSJ-QB/QBRT	18.2	20.88
(4) WSJR/WSJT	85.63	85.54
(5) WSJR/WSJT-R	85.87	83.75
(6) WSJR/QBT	44.65	48.75
(7) WSJR-QBR/WSJT	85.59	85.19
(8) WSJR-QBR/WSJT-R	86.45	84.45

[a]We note that here both parsers do somewhat better on the mirror-image WSJT data than on the standard WSJT data when trained on QB, where one might expect the opposite result, but this difference may due to the sparse nature of the standard test section

interferes in some way with the parsing of normal based sentences. Therefore, we will in general put to one side comparisons based on just this test data set, e.g., contrasts like WSJ/WSJT vs. WSJ-QB/WSJT. We also leave for future research the measurement of statistical significance of the scores by means such as stratified shuffling, as in [3], or the assessment of oracle-type scores.

The key finding to take away from these results is that there is strong evidence that both parsers were able to learn the mirror-reversal question constructions quite well, though the lexicalized BC-M2 parser was less successful. To see this result most clearly one need only focus on the histogram bar marked with an arrow

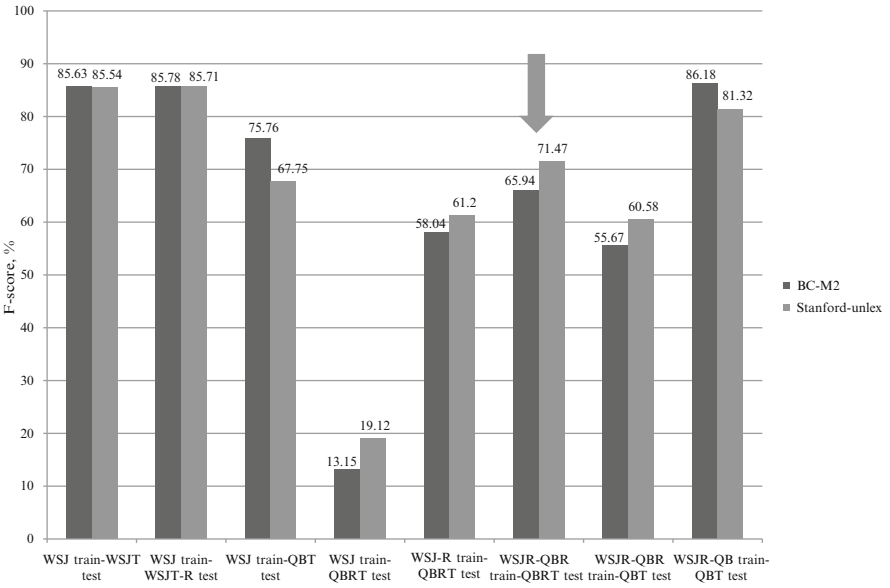

Fig. 13 F-score comparisons for BC-M2 and Stanford-unlex parsers show that the parsers do not perform well on mirror-image questions (the fourth, *middle* histogram pair from the *left*), but performance increased dramatically given QB mirror image question training, by 50 % points or more, as shown by the next two histogram pairs to the *right*. The right-most histogram repeats the finding from Sect. 3 showing that normal question parsing is also improved by the addition of normal QuestionBank training data

in Fig. 12, and note its performance gain compared to the preceding two bars, which summarize the before/after training effect. For example, when trained on only normal data, the Stanford unlexicalized parser scored only 19.12 % on the QuestionBank mirror-reversed test set, combination WSJ/QBRT, line 5 in Table 7 and the fourth histogram from the left in the figure. This number, then, may be taken as the "baseline" for a parser that has not learned anything about mirror-image questions. We may contrast this performance with training on just the WSJ reversed questions (which constitute only a small fraction, just few hundred examples out of nearly 40,000 sentences), line WSJR/QBR in the table. The initial 19.12 % figure goes up 50 % points, to 61.20 %, and additional QB mirror training examples boost this even further, another 10 % points, to 71.47 %, line 7, WSJR-QBR/QBR. Note that this is even better than the parser's performance on wh-questions after training on ordinary wh-questions. These are huge differences.

The performance gains for BC-M2 are nearly as good, though the actual numbers are less because the built-in English head-finding rules, which bias the formation of right-branching structures, cut against the grain of the mirror-reversed questions. Nevertheless, BC-M2 still performs remarkably well, as attested by examples like

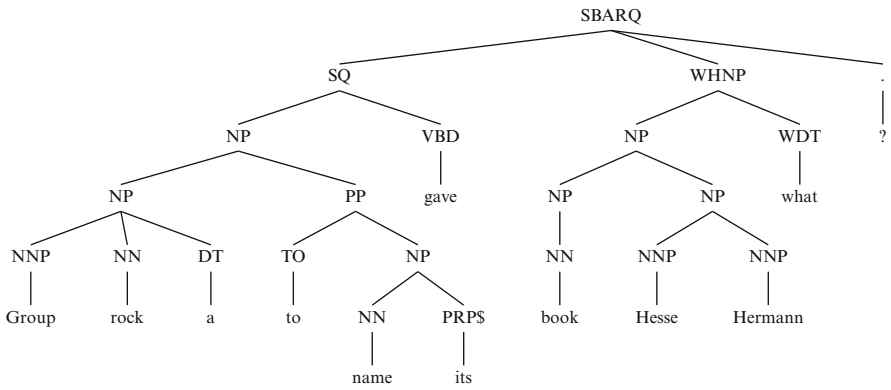

Fig. 14 BC-M2 correct parse of a "mirror" sentence from QuestionBank

the one shown in Fig. 14, the reversal of the QuestionBank sentence *What Herman Hesse book gave its name to a rock group*. Errors arise because the head rules attempt to locate heads at the left edge of phrases, except in Noun Phrases, but this of course is exactly opposite to what is required for mirror-reversed questions. A more careful experiment would re-do the BC-M2 head rules to locate heads at the right periphery, but one could then argue that we are in some sense aiding the parser in its discovery of the proper form for mirror-reversed questions. In a sense, it is startling that the BC-M2 parser works so well in spite of this handicap. Without any exposure to mirror-reversed questions, BC-M2 starts from a baseline of 13.15 %. This score rises to 58.04 %, line 6 in Table 7, a jump comparable to that of Stanford-unlex of more that 50 % in performance, after training on WSJ-TR examples. As with Stanford-unlex, training on reversed QuestionBank increases performance even further, to 65.94 % (line 7 in the table).

Row (7) and the next-to-last histogram bars in Fig. 13 the also indicate that the system has learned that questions are mirror-reversed: parsing performance drops by over 10 % when the systems are trained on WSJR-QBR, and then tested on normal questions, QBT. In short, there is every indication that mirror-image questions are learned with some facility.

It seems apparent that the BC-M2 parser could be further improved if the English-biased head-finding rules were re-written (though at the cost of "building-in" this linguistic knowledge). Figure 15 displays an example of a reversed sentence from QuestionBank, *What melts in your mouth not in your hands*, where the reversal, *Hands your in not mouth your in melts what* is given a (slightly) incorrect parse where a PP is mis-labeled as an NP. We will leave this more detailed analysis for future work.

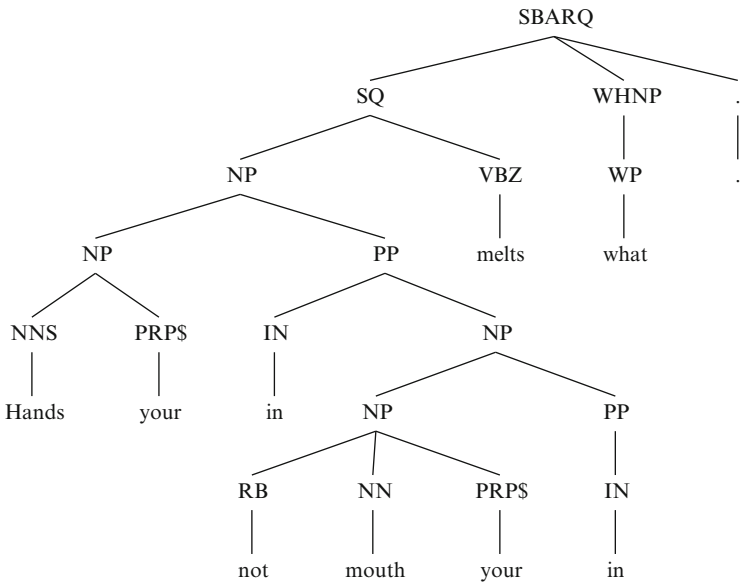

Fig. 15 BC-M2 parse of a "mirror" reversed question from QuestionBank with an erroneous labeling of a PP as an NP

7 Discussion and Conclusions

Let us now revisit the basic question outlined earlier and take stock of the results: Have state-of-the-art statistical parsers attained "knowledge of language"?

Current state-of-the-art systems, such as the several parser reviewed in this paper, score close to the 90 %-level (on withheld PTB data) when evaluated on phrase structure bracketing fidelity [16]. Of course, bracketing is not the only possible evaluation metric, as is now widely understood. In many cases, dependency relations may be of more importance; see [13] among many others for a discussion of this matter.

To the extent that such bracketing reflects linguistic knowledge, then such parsers do, of course successfully acquire that knowledge. Moreover, as noted by Petrov et al. [42] among others, modern statistical parsers can acquire tacit information about the details of verb subcategories, along with derivational structure. However, merely being able to bracket sentences "accurately" evidently does not constitute full "knowledge of language." Rather, knowledge of language is multi-dimensional and cannot be conveniently summarized in terms of a single number, an F-measure. Similarly, grammaticality cannot be described in terms of a simple probability score. We could not predict the outcome of the *read* experiment in advance simply by looking at aggregate F-measures, nor any other proposed measures we are aware of. Such conclusions may seem obvious from the outset, but the goal in applying

the kinds of stress tests described in this chapter is to discover exactly where these systems fail.

The *read* sentences are also good exemplars of such a diagnostic aid. In this case, they point to a general issue with "long distance" agreement in tense (and other features) that is not to the best of our knowledge explicitly encoded in any of the statistical models, but only indirectly, perhaps through the use of extended horizontal and vertical domains of Markovization (as in the Stanford parsers), or through the use of latent variables. Even so, as we saw in the examples of the Berkeley and CJ systems with *read*, the use of tacit, indirectly formed categories may not precisely capture the right information. Rather, the results here suggest that it may be useful to explicitly import such machinery, as is done, for example, in the statistically-grounded versions of Lexical-Functional Grammar (see, for example, [43]; unfortunately, this system is not public and was not available to us for testing).

A second unsurprising result is that many of the limitations of current systems are due to the obvious sparsity of the PTB corpus. This effect is quite clearly displayed in the relatively poor performance on wh-questions, as well as how much that performance may be boosted by simply adding new wh-questions, sometimes only a handful, as the Stanford parser example illustrates.

In this chapter we have been able to select only one or two examples out of a long list of grammatical generalizations that linguists have accumulated over the past 60 years. It remains to analyze the remainder. The challenge for future research is whether these or similar diagnostics can be exploited to advance the state-of-the-art in statistical parsing. Given such a list, and given current statistical parsing methods based on discriminative methods, it may even be possible to construct a list of both positive and negative exemplars, as with minimally different wh-question examples, and then apply the method of "contrastive estimation" developed by Smith and Eisner [50] which compares positive training examples against negative examples in the local neighborhood of the training data. Some means of "discouraging" the leap to implausible or impossible word order patterns could be a welcome side-effect of this minimal use of negative examples, eliminating the ability to infer unnatural mirror-image structure.

The pilot experiment in Sect. 5.1 demonstrates that statistically significant improvements in parsing can be achieved by regularizing passive argument structure. However, in some cases passive regularization also led to worse performance. A more careful, case-by-case analysis of these examples would seem warranted. It appears from a superficial examination of the examples where parsing performance degrades that in each instance the regularization method has partly failed, sometimes introducing additional complex structure. If so, then further improvement may be possible if one can more accurately reconstruct the underlying form, either for small clauses or for relative clauses.

It seems clear that one could apply the notion of regularization more broadly to other types of displacements, such as topicalization and dislocation structures. We predict that these will provide additional parsing improvements, possibly approaching the levels achievable only through parse re-ranking. More generally, we note

that the use of paired surface and underlying structures may provide great power not only in improving parsing, but also for providing a means to learn new rules to span the space of grammatical forms that have never been seen in training data, a major roadblock in state-of-the-art statistical systems. This is because our regularization approach bears important parallels to one of the few complete, mathematically established learnability results for a complete grammatical theory, that by Wexler and Culicover [53]. The Wexler and Culicover approach is based on a similar idea: the learner is assumed to be able to reconstruct the underlying "D-structure" corresponding to surface sentences, and from this pairing, hypothesize a possible mapping between the two. It remains for future research to determine whether this can be done for other displaced phrases in the PTB more generally.

Finally, we also note that in more recent grammatical theories, argument structure is regularized to an even greater degree by means of a VP-vP "shell structure" of branching nodes, that place Subject and then the Direct Object and Indirect Object NPs in specific, fixed positions with respect to the verb, perhaps in all languages [21]. If this is true, we could readily expand our regularization approach to this notion, which might provide a statistically-based, machine learning system with additional standardized patterns that are more easily learnable from training data alone. A full-blown incorporation of this kind of grammatical structure again remains for future work, but gives some hint at the untapped power of linguistic theory ready to be applied to treebank parsing.

Acknowledgements We would like to thank Michael Coen and Ali Mohammed for assistance and valuable suggestions. More importantly, we would like to extend special thanks to those individuals who have graciously made their parsing systems publicly available for open experimentation, in particular Daniel Bikel and Michael Collins; John Judge for his extremely valuable QBank resource and his generosity in providing it to us; Mark Johnson and Eugene Charniak; the members of the Stanford NLP group, including Daniel Klein and Christopher Manning; the Berkeley NLP group, including Stan Petrov and Daniel Klein; and the Malt and C&C parser developers. Without their generosity, analyses like those carried out here would be impossible. Finally, we would like to acknowledge two anonymous reviewers whose suggestions greatly improved this work.

References

1. Abney, S. (1996). Statistical methods and linguistics. In J. Klavans, & P. Resnik (Eds.), *The balancing act: Combining symbolic and statistical approaches to language* (pp. 1–26). Cambridge/Massachusetts: MIT Press.
2. Berwick, R. C., & Weinberg, A. S. (1982). *The grammatical basis of linguistic performance*. Cambridge: MIT Press.
3. Bikel, D. (2004a). *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. Ph.D. Thesis, University of Pennsylvania, Department of Computer Science.
4. Bikel, D. M. (2004b). Intricacies of Collins' parsing model. *Computational Linguistics, 30*(4), 479–511.
5. Charniak, E. (2000). A maximum-entropy inspired parser. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computtaional Linguistics* (pp. 132–139), Seattle. Association for Computational Linguistics.

6. Charniak, E., & Johnson, M. (2005). Coarse to fine *n*-best parser and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (pp. 173–180), Ann Arbor. East Stroudsburg: Association for Computational Linguistics.
7. Chiang, D., & Bikel, D. M. (2002). Recovering latent information in treebanks. In *Proceedings of the 19th International Conference on Computational Linguistics* (pp. 183–189), Howard International, Tapei.
8. Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.
9. Chomsky, N. (1968). *Language and mind*. New York: Harcourt-Brace.
10. Chomsky, C. (1969). *The acquisition of syntax in children from 5 to 10*. Cambridge: MIT Press.
11. Clark, S., & Curran, J. (2007). Wide-coverage efficient statistical parsing with ccg and log-linear models. *Journal of the Association for Computational Linguistics, 33*, 493–452.
12. Clark, A., & Lappin, S. (2009). Another look at indirect negative evidence. In *Proceedings of the EACL 2009 Workshop on Cognitive Aspects of Computational Language Acquisition* (pp. 26–33), Athens. Association for Computational Linguistics.
13. Clegg, A. B. (2008). *Computational-linguistic approaches to biomedical text sining*. Ph.D. thesis, Birbeck College, University of London.
14. Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (pp. 16–23), Madrid. Association for Computational Linguistics.
15. Collins, M. (1999). *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.
16. Collins, M. (2003). Head-driven statistical models for natural language parsing. *Computational Linguistics, 29*(4), 589–637.
17. Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63*, 522–543.
18. Curran, J., Clark, S., & Bos, J. (2007). Linguistically motivated large-scale nlp with C&C and boxer. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* (pp. 33–36), Prague, Czech Republic: Association for Computational Linguistics.
19. Eisner, J. (2001). *Smoothing a probabilistic Lexicon via syntactic transformations*. Ph.D. thesis, University of Pennsylvania.
20. Gleitman, L., Gleitman, H., & Shipley, E. (1972). The emergence of the child as grammarian. *Cognition, 1*(2–3), 137–164.
21. Hale, K., & Keyser, S. (1993). On argument structure and the lexical representation of syntactic relations. In K. Hale, & S. Keyser (Eds.), *The view from building 20* (pp. 53–110). Cambridge: MIT Press.
22. Hockenmaier, J. (2003a). *Data and Models for Statistical Parsing with Combinatory Categorial Grammar*. Doctoral Dissertation, University of Edinburgh.
23. Hockenmaier, J. (2003b). Parsing with generative models of predicate-argument structure. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 359–366), Sapporo, Japan: Association for Computational Linguistics.
24. Jackendoff, R. (1999). *Why can't computers use English?* New York: Linguistic Society of America (LSA) Publications.
25. Johnson, M. (1998). Pcfg models of linguistic tree representations. *Computational Linguistics, 24*(4), 613–632.
26. Judge, J., Cahill, A., & van Genabith, J. (2006). Questionbank: Creating a corpus of parse-annotated questions. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 497–504), Sydney, Australia: Association for Computational Linguistics.
27. Klein, D., & Manning, C. (2003a). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430), Sapporo. East Stroudsburg: Association for Computational Linguistics.

28. Klein, D., & Manning, C. (2003b). Fast exact inference with a factored model for natural language parsing. In *Advances in Neural Information Processing Systems* (pp. 3–10), Cambridge.
29. Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics, 43*(2), 393–427.
30. Levy, R. (2006). *Probabilistic models of word order and syntactic discontinuity*. Ph.D. thesis, Stanford University.
31. Levy, R., & Andrew, G. (2006). Tregex and tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa.
32. Levy, R., & Manning, C. D. (2004). Deep dependencies from context-free statistical parsers: Correcting the surface dependency approximation. In *Procedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (pp. 327–334). East Stroudsburg: Association for Computational Linguistics.
33. Marcus, G. (2003). *The algebraic mind*. Cambridge: MIT Press.
34. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of english: The penn treebank. *Computational Linguistics, 19*(2), 313–330.
35. Morgan, J., Meier, R., & Newport, E. (2004). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language, 28*(3), 360–374.
36. Musso, M., Moro, A., Glauche, V., Rijntjes, M., Reichenbach, J., Buchel, C., & Weiller, C. (2003). Broca's area and the language instinct. *Nature Neuroscience, 6*, 774–81.
37. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., & Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering, 13*(2), 95–135.
38. Nivre, J., Rimell, L., MacDonald, R., & Rodriguez, C. G. (2010). Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing. International Association for Computational Linguistics.
39. Parisse, C. (2012). Rethinking the syntactic burst in young children. In A. Alishahi, T. Poibeau, A. Korhonen, & A. Villavicencio (Eds.), *Cognitive aspects of computational language acquisition*. New York: Springer.
40. Petrov, S., & Klein, D. (2007). Learning and inference for hierarchically split PCFG's. In *AAAI 2007 Nectar Track*, Washington. AAAI.
41. Petrov, S., & Klein, D. (2008). Sparse multi-scale grammars for discrimininative latent variable parsing. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 867–876), Honolulu. Association for Computational Linguistics.
42. Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 433–440), Sydney, Australia: Association for Computational Linguistics.
43. Riezler, S., King, T. H., Kaplan, R. M., Crouch, R., Maxwell, J. T. I., & Johnson, M. (2002). Parsing the wall street journal using a lexical-functional grammar and discriminative estimation techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)* (pp. 271–278), Philadelphia, PA: Association for Computational Linguistics.
44. Rimmell, L., Clark, S., & Steedman, M. (2009). Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Meeting on Empirical Methods on Natural Language Processing* (pp. 813–821), Singapore: Association for Computational Linguistics.
45. Saffran, J., & Newport, E. (2007). Statistical learning in 8-month old infants. *Science, 274*(5294), 1926–1928.
46. Sekine, S., & Collins, M. (2008). The evalb program.
47. Shipley, E., Smith, C., & Gleitman, L. (1969). A study in the acquisition of language: Free responses to commands. *Language, 45*, 322–343.
48. Smith, N., & Johnson, M. (2007). Weighted and context-free grammars are equally expressive. *Computational Linguistics, 33*(4), 477–491.

49. Smith, N., Tsimpl, I. -M., & Ouhalla, J. (1993). Learning the impossible: The acquisition of possible and impossible languages by a polyglot savant. *Lingua, 91*, 279–347.
50. Smith, N. A., & Eisner, J. (2005). Guiding unsupervised grammar induction using contrastive estimation. In *International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Grammatical Inference Applications* (pp. 73–82), Edinburgh, Scotland: Association for Computational Linguistics.
51. Tateisi, Y., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). Syntax annotation for the Genia corpus. In *Proceedings of the International Joint Conference on Natural Language Processing* (pp. 222–227), JJeju Island, Korea: Association for Computational Linguistics.
52. Turian, J., & Melamed, I. D. (2006). Advances in discriminative parsing. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics* (pp. 873–880), Sydney, Australia: Association for Computational Linguistics.
53. Wexler, K., & Culicover, P. (1983). *Formal principles of language acquisition*. Cambridge: MIT Press.

Rethinking the Syntactic Burst in Young Children

Christophe Parisse

Abstract Recent proposals about children's first language acquisition have stressed usage-based acquisition and suggest that children have no language specific innate knowledge but instead use general cognitive abilities, such as perception, memory, and analogical processing, to acquire their mother tongue. These proposals do not, however, account for one argument raised by proponents of innate grammar approaches, which is the speed and correctness of children's language acquisition, which could be described as a syntactic burst usually occurring around age two to three. In this chapter, a testing procedure is proposed to demonstrate that the acquisition of usage-based and fixed-form patterns can account for this syntactic burst. The analysis is conducted with the large Manchester corpus from the CHILDES database. It is demonstrated that fixed-form patterns extracted from child input and used in their raw and unprocessed form can, if combined freely, account for the children's subsequent language production. Results show that young children's grammatical abilities (before age three) could result from simple mechanisms and that complex linguistic mastery does not need to be available early in the course of language development.

1 Introduction

Young children acquire their mother tongue very easily and without apparent effort [19]. Between age two and three, most children go through what could be called a syntactic burst. In other words, they progress from uttering one word at a time to constructing utterances with a mean length of more than three words, with frequent longer utterances, and they do this without any negative evidence and with limited

C. Parisse (✉)
MoDyCo-INSERM, CNRS, Paris Ouest Nanterre La Défense University,
92001 Nanterre cedex, France
e-mail: cparisse@u-paris10.fr

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_7, © Springer-Verlag Berlin Heidelberg 2013

input data [20]. For about 40 years, ever since Chomsky's proposals [3] concerning language acquisition and the nature of human linguistic knowledge, leading theories (generative grammar) about language acquisition postulated the existence of innate constraints on the grammar of human languages and the human mind [18,26]. In the last 10 years, a different approach to language acquisition has been advocated, which postulates completely different theoretical principles and a different interpretation of developmental language data. This approach, which can be called 'usage-based language acquisition', is rooted in 'cognitive grammar theory', and especially construction grammar theory [6]. Construction grammar postulates that linguistic knowledge is based on constructions, form-function pairings, which may be highly idiosyncratic and correspond to words or multi-word fixed-forms, or may be general and correspond to what is referred to as rules in generative grammar. Cognitive grammar has a different stance from generative grammar towards the nature of linguistic knowledge, in that item specific knowledge is considered as the most frequent situation and generalized or semi-generalized patterns as the less frequent situation (cf. discussion in [6], Chap. 3). Recent work on language acquisition [23, 24] showed that children did not in fact demonstrate early knowledge of general categories such as noun or verb, but that on the contrary grammatical knowledge is built up in a piecemeal fashion. Children first learn to generalize fixed constructions around specific items. More general knowledge is developed only slowly and, for example, general knowledge of the verb category appears only around age four [23], whereas knowledge of the noun category appears at about age two.

2 Assumptions About Children's Behavior

The goal of the current paper is to demonstrate that it is possible to account for the syntactic burst on the basis of the sole use of fixed-form patterns extracted from input by children. The demonstration is based on two assumptions about young children's perceptive and mnemonic capacities: anything they have once produced, they can produce again; and, when their language exactly reproduces an adult's, this can be explained as a simple copy of their input. Nothing is assumed about the length of the elements copied by the children (see [15]). These elements can correspond to one or more adult target words, or even to word parts. For example, one element could be 'juice' or 'of juice', another could be 'drink' or the more complex 'little drink'. These two assumptions do not imply the existence of specific grammatical knowledge but rather of general cognitive abilities (especially auditory pattern extraction and long term memory [4, 7, 8]). A third assumption is that children can produce and combine these fixed-forms at will using simple concatenation (i.e. the production of several patterns in succession in a single prosodic phrase—see Konopczynski [10]). For example, one combination could be 'drink juice' and another 'drink of juice'. This allows children to produce new, longer or more complex utterances than what they hear from their input. The concatenation of

patterns by young children in a single prosodic forms is attested [7, 10], but it is yet unknown when and how this concatenation is organized. It could be random, organized by non grammatical principles, or organized by grammatical principles (see Analysis 3 below). We therefore make no assumptions concerning this process.

3 A Testing Procedure in Three Steps

The hypotheses presented above were tested using a corpus of adult and child spontaneous language interaction. The testing procedure was divided into three steps. The goal of the first two steps was to identify the fixed-forms that children use and to check whether these patterns could indeed be extracted from the children's input. These two successive steps create, in an iterative process, a list of fixed-forms that were used for the final step of the testing algorithm. The goal of this final step was to check whether utterances produced by the children which were not made of a single fixed-form could indeed be produced by the concatenation of several fixed-forms identified in the first two steps. Identifying the fixed-forms used by the children presents a technical challenge: when an utterance produced by a child contains more than one adult target word, how is it possible to know whether this pattern is extracted as a fixed-form from the input or whether this pattern is constructed by the child? To solve this issue, we chose to start with patterns that could not be constructed by the child, but only copied from the input, and we built our analysis on the basis of these patterns. These patterns, that were considered as always extracted from the input and not further processed by the children, were the children's utterances that corresponded to a single adult word. The identification of these utterances was the goal of the first step of the testing procedure. For example, if the child says 'jump', this word is added to the child's basic patterns. If 'jumps' also occurs, this word is also added to the basic patterns, as we make no assumption about the child's ability to decompose words into morphemes (to add an 's' to 'jump'). All these isolated words were entered in the 'list of fixed-forms'. The goal of the second step was to analyze children's productions containing more than one adult target word. When this was the case, the utterance was analyzed to check whether it contained one or more of the previous fixed-forms. If it contained only one or no previous fixed-form, then this utterance was added to the fixed-forms. For example, 'a jump' is added because 'a' was not a previous fixed-form and 'jump' was. This means that the list of fixed-forms can include elements made of more than one word, but these elements are considered as 'one fixed-form', they are never decomposed. If the child's production contained two or more fixed-forms, they were considered as made of more than one piece, and were used for testing the children's productivity. For example, 'I jump a kangaroo' contains 'jump' and 'kangaroo' that are produced as fixed-forms by the child, so it is an utterance that is not considered known as a fixed-form but constructed out of previous material. This procedure is iterative. For each utterance under scrutiny, if it is a single word then step 1 applies, and if it

contains one or zero fixed-forms, then step 2 applies. This continuously increments the list of fixed-forms. If it contains more than one fixed-form, the testing list is incremented. For the next utterance, the new version of the list of fixed-forms is used, which simulates the growing knowledge of the children. The goal of the final and third step of the testing procedure is to measure whether the utterances produced by the children and that are not fixed-forms can be recomposed using only fixed-forms.

- Step 1: All single-word utterances produced by children are meaningful to them; they are directly derived from adults' output. They are the basic elements that children use to build language.
- Step 2: Children's multi-word utterances containing only one word already produced in isolation (words produced in step 1), along with other words never produced in isolation (never produced at step 1), are also basic elements that children use to speak. They are also directly derived from children's input; this is facilitated by the children's knowledge of isolated words. These multi-word utterances are manipulated and understood by children as single blocks, just as isolated words are. They may also be called frozen forms.
- Step 3: Children link utterances produced at steps 1 and 2 to produce multi-word utterances with more than one word already produced in isolation (words produced in step 1). They do this using a simple concatenation mechanism and the fact that the utterances they create have a pertinent meaning prevents them from producing aberrant utterances. The goal of the third step is to check whether the basic elements identified in step 1 and 2 are sufficient to account for the children's multiword utterances.

Since the productions of children and their adult partners are easy to record, it is possible to test whether the testing procedure has sufficient generative power to account for all children's productions. However, such a demonstration may be more difficult than it appears, for several reasons. First of all, the assumption made in step 1 is not always true, as it is quite possible for a child to reproduce any sequence of sounds while playing with language. This uncertainty about step 1 is only important in conjunction with step 2, as isolated words are the key used to parse the elements of step 2. To decide that a word has meaning in isolation for a child, it has been assumed that it must first have meaning in isolation for an adult. Words in the categories of determiner and auxiliary produced in isolation have been considered as not having meaning in isolation and have therefore been removed from the elements collected in step 1. Analysis of language data demonstrated that this assumption is quite reasonable, as the use of these words in isolation is often the result of unfinished utterances, with incomplete prosody. Measuring the generative power of the testing procedure implies evaluating the accuracy of the assumptions made in steps 1, 2 and 3. It is easy to imagine that these assumptions hold for the first multiword utterances of young children, before age two. The question is: to what extent is this true and until what age? Four analyses have been carried out in order to answer this question.

4 Analysis 1

The first analysis uses the testing procedure exactly as it is described in its principle above. The analysis is based on the Manchester corpus [22] from the CHILDES database [13]. The corpus contains recordings of 12 children from the age of 1;10–2;9. The mean length of utterance in words varies from 1.5 to 2.9. Each child was seen 34 times and each recording lasted 1 h. This results in a total production of 537,811 words in token and 7,840 in type. For each child, the average is 44,817 words in token (SD = 9.653) and 1,913 in type (SD = 372). The testing procedure was run iteratively in three steps. Each step from the procedure corresponds to one of the parts described above.

- Step 1: For each transcript, the child's single-word utterances are extracted and added to a cumulative list of words uttered in isolation, referred to as L1. It is possible to measure at this point whether the words on L1 can be derived from the adult's output. In order to do this, a cumulative list, L-adult, of all adult utterances is also maintained.
- Step 2: For each multi-word utterance in the transcript, the number of words previously uttered in isolation is computed using list L1. Multi-word utterances with only one word uttered in isolation are added to a list called L2. It is possible to measure at this point whether the utterances on L2 can be derived from the adult's output (list L-adult above).
- Step 3: the remaining utterances (list L3), which contain more than one word previously uttered in isolation, are used to test the final step of the algorithm. The test consists in trying to reconstruct these utterances using a catenation of the utterances from lists L1 and L2 only. Two measurements can be obtained: the percentage of utterances on list L3 that can be fully reconstructed (referred to below as the 'percentage of exact reconstruction') and the percentage of words in the utterances on list L3 that contribute to a reconstruction (referred to below as the 'percentage of reconstruction coverage). For example, for the utterance 'The boy has gone to school', if L1 and L2 contain 'the boy' and 'has gone' but not 'to school', only 'the boy has gone' can be reconstructed, thus leading to a percentage of reconstruction coverage of 66 %. The percentage of exact reconstruction is the percentage of utterances with a 100 % reconstruction coverage. The percentages of list L3 that are reconstructed or recovered do not include utterances from L1 and L2 lists.

The testing procedure is iterative because it is performed in turn for each of the transcripts of the corpus. Lists L1, L2 and L-adult are cumulative, which means that the lists obtained with transcript 1 are used as a starting point for the analysis of transcript 2, and so on. This presupposes that children can reuse data they heard only once a long time after they heard it. In Step 1 the percentage of words in L1 present in adult speech has a mean value of 91 % (SD = 0.03). In Step 2, the percentage of elements of L2 present in adult speech has a mean value of 67 % (SD = 0.05). These two results are stable across ages—even though lists L1, L2 and

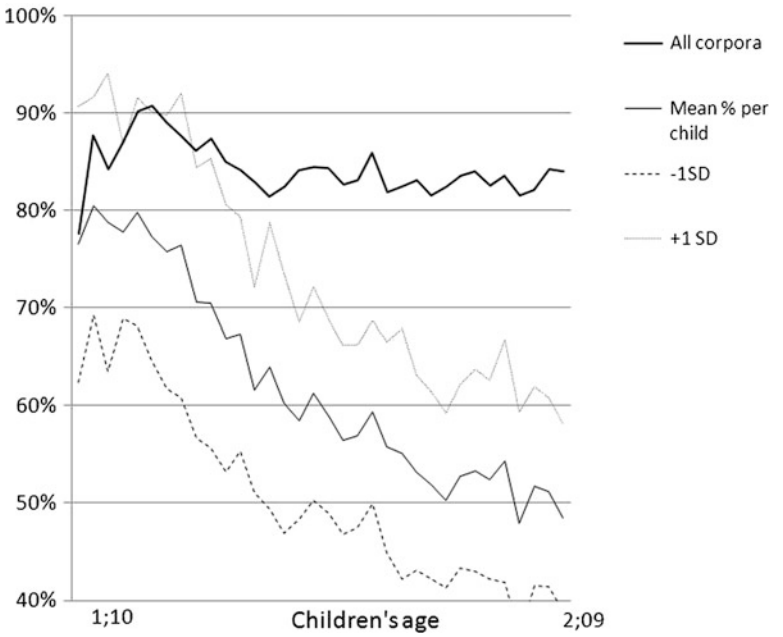

Fig. 1 Percentage of utterances exactly reconstructed

L-adult grow continuously. After two transcripts, for all 12 children, lists L1 + L2 represent 11,979 words in token and L-adult contains 82,255 words in token. After 17 transcripts, these totals are 89,479 and 688,802, respectively. After 34 transcripts, they total 167,149 and 1,370,565. The ratio comparing the size of L1 + L2 and L-adult does not evolve much, varying between 6 and 8.

The results for Step 3 are presented in Figs. 1 and 2. Each point in the series corresponds to the nth iteration performed with the nth transcript. The mean value is the mean of the percentage for all children considered as individuals (reconstruction between a child's corpus and his/her parents' corpus only). The algorithm was also applied to all corpora: for each time point in the series of recordings, the 12 files corresponding to 12 children were combined into a single file used to run the nth iteration of the algorithm. Percentages for all corpora are shown with a bold line in Figs. 1 and 2. The percentages are clearly higher for the aggregated corpora. The percentage of words in L1 present in adult speech goes from 91 % up to 97 % (SD = 0.01). For L2, the percentage goes from 67 % up to 80 % (SD = 0.07). The percentage of exact reconstruction goes from 62 % (SD = 10 %) up to 84 % (SD = 3 %) and the percentage of reconstruction coverage goes from 87 % (SD = 3 %) up to 95 % (SD = 1 %). The number of unknown utterances (list L3) increases more than the number of known utterances (lists L1 and L2). After two transcripts, there are half as many elements in list L3 as in L1 + L2. But after 17 transcripts, L3 is 42 % larger than L1 + L2, and after 34 transcripts, it

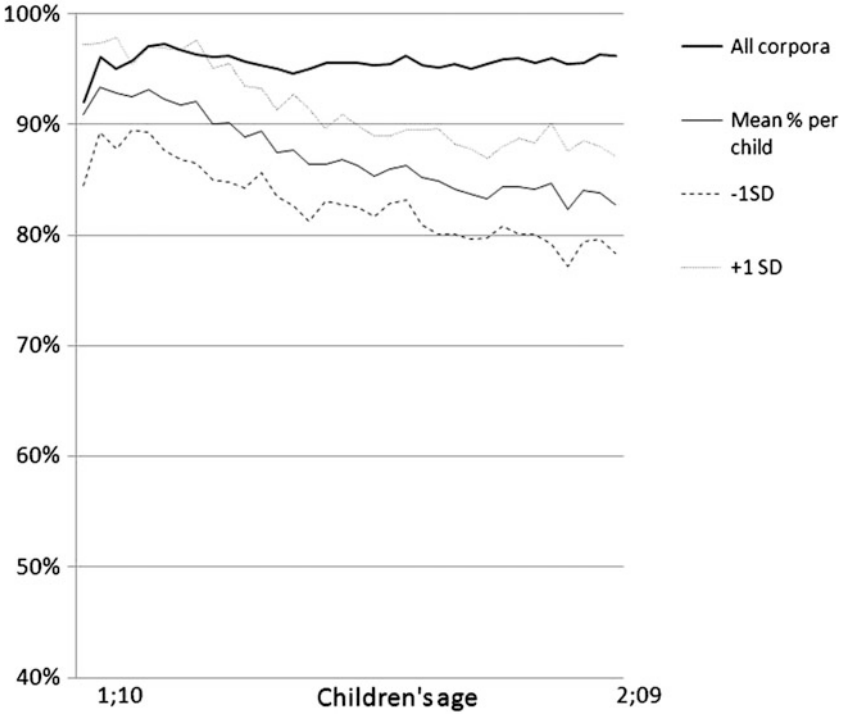

Fig. 2 Percentage of reconstruction coverage in all utterances

is 127 % larger. As children grow older, there is a decrease in the scores for exact reconstruction and reconstruction coverage. This decrease is greater in individuals than for the children as a group, which suggests a size effect.

5 Analysis 2

The goal of the second analysis is to compare the results of the first analysis with a baseline and to test whether knowledge of general syntactic categories such as noun and verb would help the children. The baseline corresponds to the case when children are just reproducing what they heard verbatim, without combining elements. This means that any utterance produced by a child is stored in the child's memory (even if it is made of more than one element produced in isolation) and reused when necessary. The elements used in this analysis include lists L1, L2 and L3 from analysis 1. The current analysis uses the same corpus as the first one. The results for Step 1 and Step 2 do not change because these steps did not involve combination of stored information. The results for Step 3 are presented in Fig. 3 (for exact reconstruction) and Fig. 4 (for reconstruction

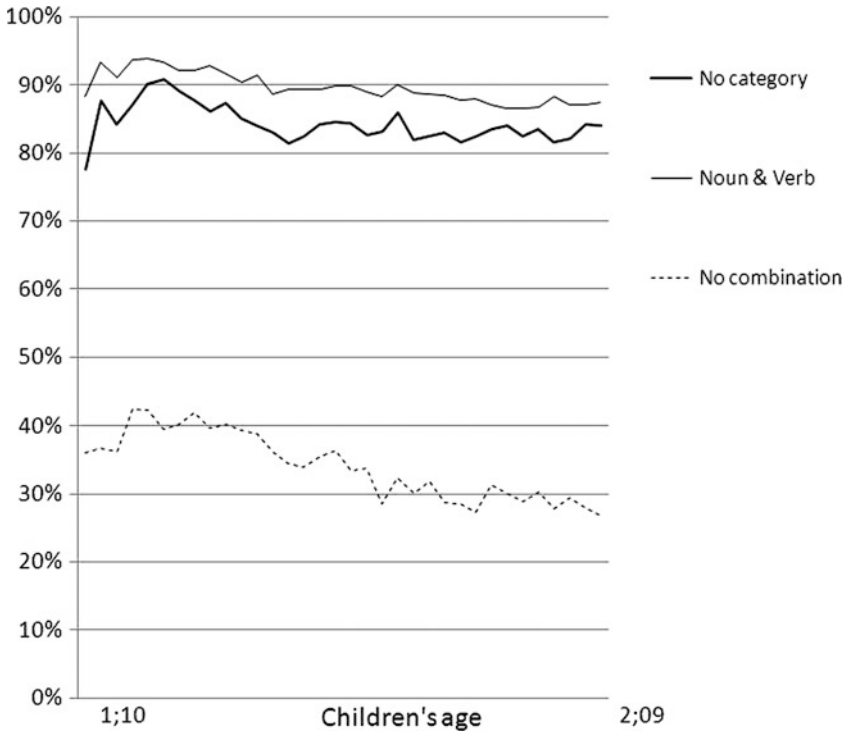

Fig. 3 Percentage of utterances exactly reconstructed, depending on the degree of knowledge of noun and verb categories

coverage). The results correspond to the 'no combination' lines in the figures. There was a clear improvement between the baseline and the previous results (see 'no category' in Figs. 3 and 4). The percentage of exact reconstruction went from 34.0 % (SD = 4.9 %) to 84.5 % (SD = 2.5 %) and the percentage of reconstruction coverage went from 67.8 % (SD = 2.8 %) to 95.6 % (SD = 0.8 %).

As the results of the first analysis do not reach 100 %, it is important to find out what type of information could increase the quality of the results. One way is to increase the size of the corpus. This was done in the first analysis by applying the testing procedure on all 12 corpora considered as a single whole. This provided a substantial improvement but did not reach 100 %. Another improvement could be obtained if we consider that children can take advantage of the knowledge of major syntactic categories such as noun and verb. In many classical approaches of language acquisition, knowledge of word categories is considered as occurring very early during language development (see [18, 19]). If children have knowledge of the syntactic categories Noun and Verb, they would be able to generalize the syntactic knowledge they have already acquired. The conditions of Step 2 and Step 3 would be more easily fulfilled if the children had a certain amount of

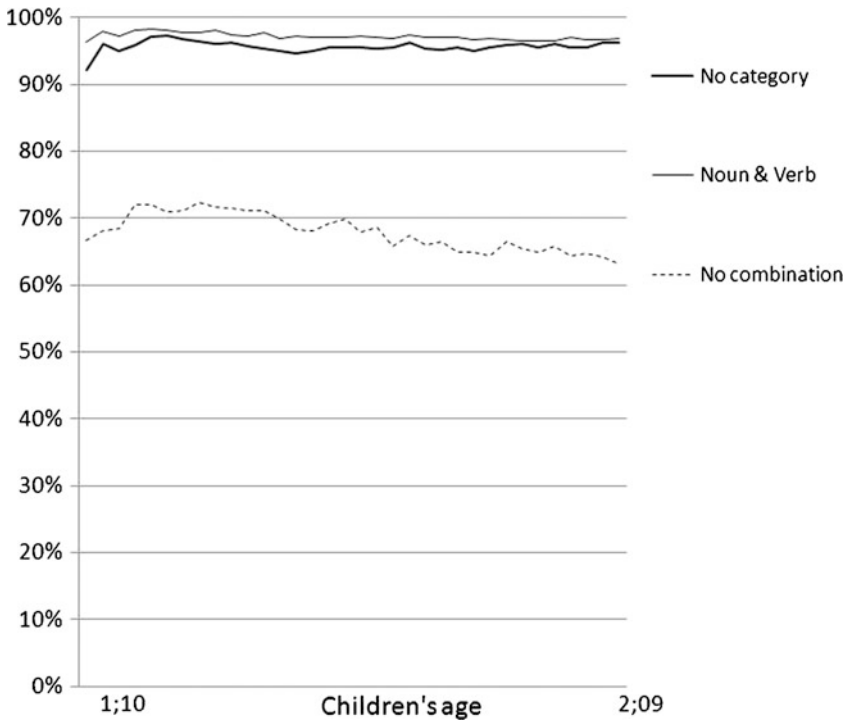

Fig. 4 Percentage of reconstruction coverage in all utterances, depending on the degree of knowledge of noun and verb categories

syntactic class knowledge. As described by Maratsos and Chalkley [14], it is possible for children to learn syntactic classes from the contexts in which words occur. However, knowledge of part of speech is unlikely in very young children on the basis of syntactic distribution. Semantic knowledge can also help to construct syntactic knowledge [1] for classes such as common nouns, proper nouns and verbs, and perhaps also adjectives and adverbs. To simulate the fact that children could construct the classes of common nouns, proper nouns and lexical verbs, every occurrence of common or proper nouns in the Manchester corpus should be substituted by the symbol 'noun' and every occurrence of non-auxiliary verbs by the symbol 'verb'. This is easy to implement because the Manchester corpus is fully tagged for part of speech, as described in the MOR section of the CHILDES manual [13]. The result is that list L1 now includes all nouns, all verbs plus all words occurring in isolation, as in the first analysis. In list L2, in utterances that include a word from the categories Noun or Verb, this word is substituted by the symbol 'noun' or 'verb'. These utterances now form rule-like productive patterns known as formulaic frames [16] or slot-and-frame structures [11]—for example, 'my + NOUN'. Another interpretation of this analysis is to consider the extension of all patterns containing a noun or a verb as a means to extend the training corpus. It

is equivalent to considering that children heard the formulaic frames for all nouns and verbs that occur in their input. Such a corpus would then correspond to a larger corpus, but with variety occurring only on nouns and verbs, not on the grammatical frames (no new frames are created with this analysis). When we reproduce the first analysis under these conditions, the new results obtained at Step 2 and Step 3 should be better, in that they should correspond more closely to the adult input, and remain constant over age. The results for Step 1 and Step 2 are indeed better than before. The percentage of utterances on L2 present in adult speech has a mean value of 91 % (SD = 0.02). The results for Step 3 are presented in Fig. 3 (for exact reconstruction) and Fig. 4 (for reconstruction coverage). In each of these figures, three results are presented for the whole Manchester corpus: one assuming no category knowledge, one corresponding to the previous results (analysis 1), and one assuming knowledge of the three categories proper noun, common noun and verb. The percentages of reconstruction become markedly higher, as any combination that contains some of the three categories proper noun, common noun and verb is known for all occurrences of the word from these categories. The mean for exact reconstruction with 'no category' knowledge is 84.5 % (SD = 2.5) and 95.6 % (SD = 0.8) for reconstruction coverage. These values increase to 89.6 % (SD = 2.2) and 97.1 % (SD = 0.5) for 'noun and verb' knowledge.

6 Analysis 3

The results obtained in the first two analyses show that the postulated mechanisms, while they do not include any innate specific linguistic mechanism, can generate the type of output produced by young children. Of course this does not demonstrate that the children did indeed use such a strategy. One way to address this issue would be to find out whether some characteristics of the children's actual production match those of the output of the procedure tested above. This issue is addressed below in two different ways. The first way is to look at the forms that children produced and that do not correspond to adult input. These forms (usually called 'errors' as they do not correspond to 'normal' adult grammar) may have specific features that are very valuable for understanding children's language development. For example, the production of 'goed' instead of 'went' shows that children are able to add a suffix ('ed') on verbal forms in certain cases. This shows the existence of the children's ability to generate past tense forms. The goal of the current analysis is to check children's errors against the usual adult norm and to find out whether these errors could come from the use of the concatenation of fixed-forms described in this paper. The second way is to find out whether children order the fixed-forms that they produce or not. If they do not, this would match the idea proposed and tested above that children have no model of fixed-form order at their early stage of language development. If they do, this would suggest that children may follow the concatenation procedure but that knowledge about fixed-form order is available to the children. More specifically, the following issues are addressed below: (1) The

characteristics of children's errors should be compatible with the properties of the testing procedure. (2) When children make word order errors, these should entail inversions of words uttered alone or blocks of words grouped around a word uttered in isolation, but not inversions of words that are never uttered in isolation. As children get older, they gradually learn to copy the adult word order, so that word order errors should occur less often than with young children.

6.1 Results and Discussion: Question 1

Given that errors with respect to the adult norm for oral English language are indicated in the Manchester corpus, it is possible to make a typology of and the errors to see if they match the properties of the testing procedure. Two formats for error descriptions exist. If some morphosyntactic or grammatical element is clearly missing, the format uses the '0' notation. This means that the missing element is transcribed, but with a 0 sign before it. For example: *CHI: what 0is [*] this. *CHI: Warren-0's [*] hair. In all other cases, the error is only signaled by a '[*]' sign. For example: *CHI: me [*] play. *CHI: foot-s [*]. There are 12,216 child errors tagged as such in the Manchester corpus. Of these, 9,063 correspond to missing elements. There are 35 different types (see Appendix, Table 2) of missing elements in the corpus, out of a total of 9,253 tokens—there may be more than one missing element per utterance. Examples of the ten most common types of error are, in order of frequency: *CHI: baby 0is [*] stuck *CHI: I 0am [*] write-ing *CHI: they 0have [*] gone *CHI: all 0are [*] eat-ing table *CHI: it 0has [*] gone *CHI: Daddy-0's [*] thumb *CHI: Andy want-0es [*] it *CHI: there two penguin-0s [*] *CHI: what-'is he do-0ing [*] *CHI: I bang-0ed [*] it The examples of each type of error have been randomly chosen from the recordings of the youngest children. In these examples, all the words produced are also used by the children in isolation or as a group in a single utterance. In particular, this is the case for 'I bang', 'what-'is', and 'he do'. The only exception is the 'I' in the second utterance 'I writing'; all the other utterances could have been produced by the testing procedure. 'I', however, is not found in isolation. As the problem raised by 'I' is also raised by 'a' in the utterances where there is no obvious grammatical element missing, we will first discuss this specific point. The 3,153 errors that do not correspond to missing grammatical elements are more diverse than the errors involving missing elements. One common type is the use of determiner 'a' with a plural noun or other inappropriate word, for example, 'a car-s', 'a flower-s', 'a apple', 'a people', 'a same'. There are 121 such errors. The problem raised by 'I' and 'a' may correspond to two different interpretations. The first is that it is by no means certain that 'I' and 'a' are never used in isolation. In fact, they are, as 'I' and 'a' may happen to be the last element of an incomplete sentence. This occurred 65 times for 'I' (46 times in isolation) and 227 times for 'a' (29 times in isolation). It may be that what is considered to be an incomplete sentence by an adult is not so from the child's point of view. If this is the case, then the testing procedure could produce utterances such as' I writing' and 'a

pants'. The second possibility is to interpret this phenomenon as the emergence of mechanisms other than the testing procedure. 'I + x' and 'a + x' are clearly very productive patterns in young children's language. There are 1,316 utterances of the type 'I + x' (with 216 different values for x) and 2,030 occurrences of the type 'a + x' (with 552 different values for x). This represents 7.4 % of all two-word utterances. These two patterns could be the first slot-and-frame structures used by children [11,17]. The testing procedure is obviously not the only mechanism used to produce language, and this may be an example of another mechanism gradually coming into play. Another type of error unaccounted for by the testing procedure is the overgeneralization of morpholexical constructions. A common example (166 occurrences) is an incorrect use of the plural marker s, for example 'milk-s', 'foot-s', 'smoke-s'. This may be explained by the same slot-and-frame mechanism as above. Most other errors are perfectly accounted for by the testing procedure. One of the most common ones is the use of words such as 'me' or 'my' as obligatory subject pronouns or existence verbs, for example, 'me play', 'me sit down', 'me egg', 'me tea', 'my make a tower', 'my do that'. This occurs 615 times for me and 210 times for 'my'. This is perfectly accounted for by the testing procedure, as 'me' and 'my' are both produced in isolation by the children. Other examples of words commonly used to build utterances in a similar fashion are 'no' (167 occurrences) and 'mine' (47 occurrences) in constructions such as 'no fit', 'no away', 'mine doggie', 'mine water'. For the purpose of this article, it is unnecessary to go through all possible types of errors. After deducting all the errors already accounted for, there are still 429 different words preceding the errors (as marked in the Manchester corpus) and 393 different words following them. It is interesting to note that many errors look as if they result from the concatenation of two elements. For example, 'mine [*] cover', 'do it [*] the animal', 'draw another one [*] fish', or 'I want [*] need my sock-s on'. It is possible to check automatically whether this is true or merely an impression. If these errors come from the simple concatenation of two strings of words, then the pair of words located around the error (in the examples above, this corresponds to the pairs 'one fish and want need') may be a creation of the child's, and thus less likely to be found in adult utterances. The other pairs of words (in the two examples above, this means the pairs 'draw another', 'another one', 'I want', 'need my', 'my sock-s' and 'sock-s on'), because they belong to strings of words extracted from children's input, would be more liable to be found in adults' utterances. All these pairs have been extracted; 1,384 different pairs located around the errors were found and 3,584 pairs located elsewhere. Of the pairs located around the errors, 674 are found in adult utterances (49 %); of those located elsewhere, 2,475 are found in adult utterances (69 %). This result confirms the plausibility of children's following the testing procedure.

6.2 Results and Discussion: Question 2

There are three different types of word inversions. The first is the inversion that occurs between isolated words or words grouped around a word used in isolation

Table 1 Percentages of inversions involving pairs of words

	All inversions	Morpholexical inversions	Group inversions
For the corpus as a whole			
any pairs	8.35	3.68	7.76
Pairs occurring twice	23.60	5.91	22.72
Transcript by transcript			
any pairs	1.75 (1.43)	0.52 (1.34)	1.59 (2.03)
Pairs occurring twice	10.78 (10.87)	2.05 (5.58)	12.54 (17.19)

Note: Standard deviations are given in parentheses

(lists L1 and L2 discussed above). For example, 'baby stuck' vs. 'stuck baby', where both words belong to list L1, or 'that one there' vs. 'there that one', where 'that one' belongs to list L2 and 'there' to list L1. The second is the inversion that occurs within words grouped around a word used in isolation (within one element of the list L2 above). For example, 'is it a baby' vs. 'it is a baby', where both groups belong to list L2. The third type occurs anywhere and between any type of words, for example 'baby a'. The words involved in this type of inversion do not belong to lists L1 or L2. There are two different ways of computing the proportion of inversions, depending on the number chosen as a reference. The first possible reference number is the number of possible word inversions. The second is the number of possible word inversions, but taking into account only the pairs of words that appear at least twice, in whatever order. The percentages of inversion can also be computed in two different ways: either for the corpus as a whole or transcript by transcript. The first option is probably felt to be more 'fair', because there is no reason why an inversion should occur during one particular recording and not during others. But the second option is the only way to show that the same child is using the same words in a variable order, within a period short enough to judge that the variability is warranted by the child's grammatical knowledge. All the values are computed in types.

The results are presented in Table 1. The results per transcript, for pairs of words that appear at least twice and sorted by age are presented in Fig. 5. The percentage of inversion is much larger for the corpus considered as a whole than for single transcripts, which is not surprising as there are many more circumstances where semantics may lead to word reversal in a large than in a small transcript. Also, pairs in any order are more frequent when one considers frequent pairs of words only. When percentages are computed transcript by transcript, it becomes possible to test the significance of the difference between the types of inversions. A t-test computed across children in the frequent pairs case shows that the difference between type 1 and type 2 is highly significant, $t(11) = 5.67$, $p < 0.00001$, as is the difference between type 2 and type 3, $t(11) = 9.40$, $p < 0.000001$. The difference between type 1 and type 3 is not significant, $t(11) = 1.55$, $p = 0.07$. Results computed across age give exactly the same pattern of results, as do statistics computed using percentages for all pairs of words. In this case, percentages are lower, but the relative ordering of results is the same. Examples of variable order between groups of words are presented in the appendix, Table 3 for the youngest children. There are two main

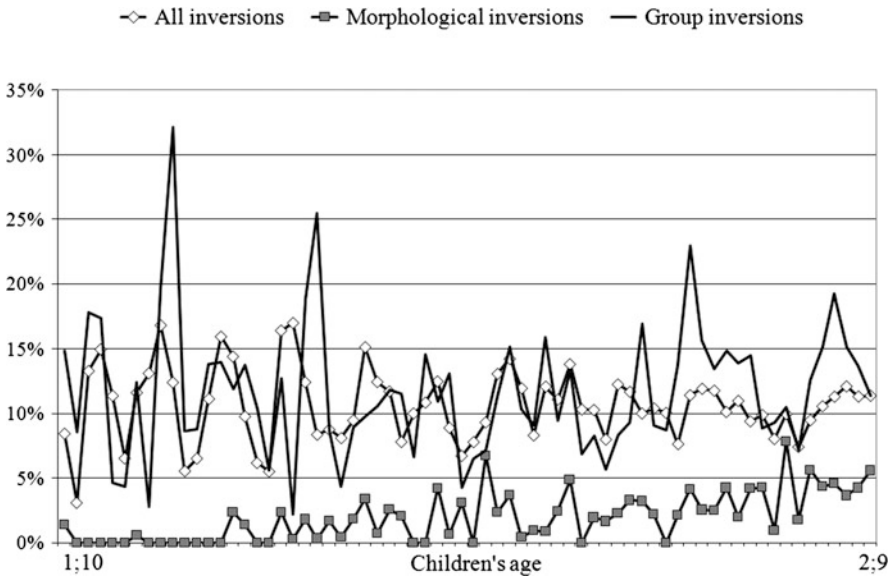

Fig. 5 Percentages of inversions for pairs found at least twice in the corpus, computed per transcript and per child

categories of variable orders. The transcript 'Warren 01A' gives a perfect representation of these two categories. The first type is 'Controller gone' vs. 'gone Controller'. Here is a pure inversion of two apparently equivalent elements. The second type is 'there brick there' which is a variable order in itself. In this case, it would seem that the two elements produced, 'there' and 'brick', have no order and that the child repeats them to emphasize what she wants to say. Inversions within a group of words from list L2 are very unusual, and only two cases seem to occur. The first is the very common pattern of pronoun and auxiliary inversion in questions ('I can' vs. 'can I', 'they are' vs. 'are they'). The second includes repetitions ('got a got a rabbit', 'in a in a minute') and coding or segmentation errors. It thus appears that inversions between a function word and a content word are impossible other than in questions. On average, the percentage of variable word order occurrences within a transcript is not very high. If the order of content word constructions were really free, the percentage of constructions in any order should have reached the 100 % level. However, for the pairs that appear twice, there is a non-negligible percentage of words that appear in two different orders, so it is difficult to decide whether word order is chosen on a morphological basis—which would imply a strict respect of word order—or on a semantic one—which would allow more laxity in word order. In any case, it is true that word order is a strong syntactic feature of the English language and that it has to appear at some point during the development of syntactic structures. The most important result here is that inversions are much more frequent between words grouped around a content word than between a functional word—that is a word that never appears in isolation—and a content word. This shows that

(1) it is when semantics has the highest content that the word order is the most free; (2) word order is only meaningful—at first—in the case of words which tend to occur together and are very frequent in the children's input (frozen forms). It is very difficult to discover a word order rule applying to two content words (such as nouns and verbs), unless either the category of these words is known or the words are very frequent. If young children follow word order more in morphological situations—the repetitive ones—than in semantic situations—which are less repetitive—this could just mean that they have not yet learned the syntactic categories and are still learning language on an example-driven basis. As for the changes in the proportion of the various word orders through time, it does not seem that our hypothesis is confirmed. As can be seen in Fig. 5, the number of variable word order elements is stable with age and only a slow decrease in variability is apparent. This would mean that word order inversions are not a developmental feature, but an intrinsic pattern of the English language, and that young children are as sensitive to word order as older children are. It could also mean that the basic characteristic of the testing procedure, that children have no model of word order—with the exception of the morphosyntactic derivation of words—holds true until at least age three.

7 Analysis 4

One limitation of analyses 1 and 2 is that nothing indicates how long the three-step mechanisms may remain efficient and appropriate. We have suggested that these mechanisms remain operational at an older age, because the use of fixed-forms and constructions is thought to be found even in adults (see [6]). This can be checked using other material from the CHILDES database with recordings spanning a longer period. The corpus chosen for this test is Brown's [2] Sarah corpus, which ranges from age 2;3 to age 5;1. The mean length of utterance in words varies from 1.47 to 4.85. This results in a total production of 99,918 words in token and 3,990 in type. In Step 1, the percentage of words on L1 present in adult speech has a mean value of 90 % (SD = 6.5). In Step 2, the percentage of elements of L2 present in adult speech has a mean value of 45 % (SD = 13.1). These two results are stable across ages. With the assumption of knowledge of the Noun and Verb categories, results for Steps 1 and 2 are, respectively, 83 % (SD = 13.8) and 55 % (SD = 16.6). The results for Step 3 are presented in Fig. 6 (for exact reconstruction) and Fig. 7 (for reconstruction coverage). In each of these figures, three results are presented: one assuming no category knowledge, one assuming knowledge of the three categories Proper Noun, Common Noun and Verb, and one baseline assuming no combination. The mean for exact reconstruction with 'no category' knowledge is 54.3 % (SD = 13.6) and 85.8 % (SD = 4.2) for reconstruction coverage. These values increase to 68.3 % (SD = 11.0) and 90.7 % (SD = 3.2) for 'Noun and Verb' knowledge. With no combination, the values decrease to 16.9 % (SD = 7.4) and 51.2 % (SD = 6.9). The average percentages of reconstruction are lower for the Sarah corpus than for the Manchester corpus. Comparing Figs. 3 and 6 and Figs. 4

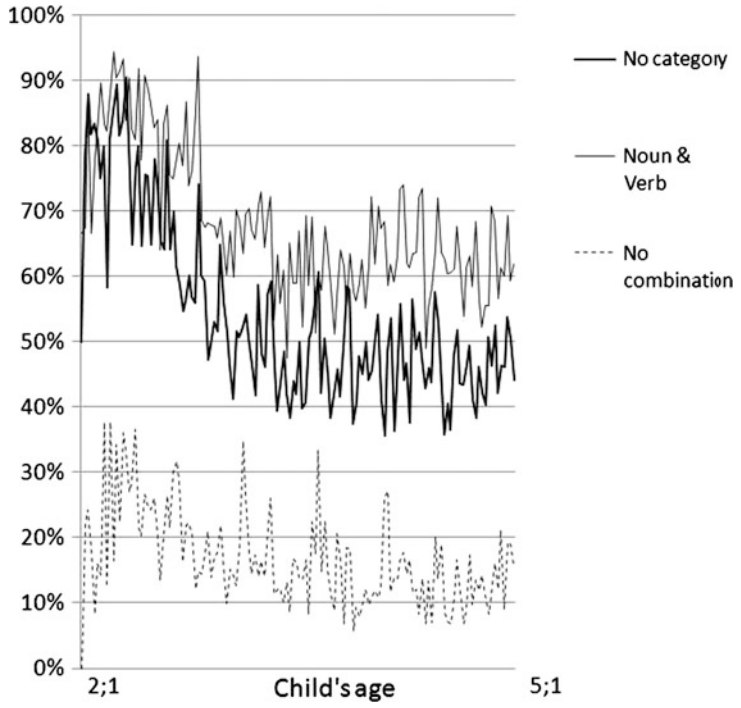

Fig. 6 Percentage of utterances in the Sarah corpus exactly reconstructed, depending on the degree of knowledge of vocabulary and syntactic categories

and 7, one can see that there is a drop in the reconstruction performances in the third year. The percentages for Sarah in her second year were as high as those for the Manchester corpus children. Part of this drop in performance may be attributed to the smaller corpus. Indeed, comparing Figs. 1 and 3 and Figs. 2 and 4, it appears that the drop in performance that became visible when single child corpora were used was not in evidence when all the corpora were amalgamated into one big corpus. It is also possible that the drop in performance found in the Sarah corpus reflects a progressive decrease in the systematic use of a simple concatenation procedure by the child. Also, the drop is marked between age 2 and age 3, but stops after about age 3;6. This could mean that, as proposed by Goldberg [6], older children as well as adults use a lot of fixed-forms.

8 Discussion

A procedure was tested on the basis that children are able to extract words or word patterns that they consider as frozen forms and that they combine using concatenation only. Such a procedure allows the children to produce language

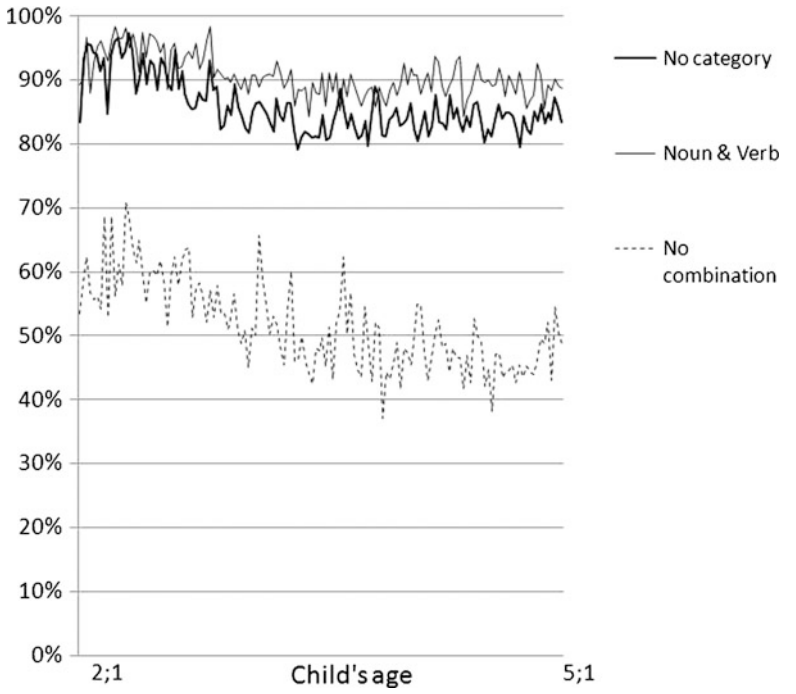

Fig. 7 Percentage of reconstruction coverage in all utterances in the Sarah corpus, depending on the degree of knowledge of vocabulary and syntactic categories

which is mostly grammatical and to generate language without having syntactic knowledge. For example, the procedure can produce the utterance 'this one here' where 'this one' and 'here' are frozen forms but 'one this' is not. So this means that whereas the production of 'here this one' is possible, the production of 'one this here' and 'here one this' is not. The procedure was tested on the Manchester corpus [22] and Brown corpus [2] from the CHILDES database [13]. The testing procedure did not achieve 100 % reconstruction in the test conditions described above, where the database consisted only of 34 one-hour recordings for each of the 12 children in the corpus. This corresponds globally to a pseudo-corpus of 408 h, which amounts to 8–10 weeks of speech. With a larger corpus, the results would probably be better, as indicated by the increase in the results when considering the corpus of a single child in comparison with the 12 times larger corpus composed by grouping all the children's data in a single corpus (see Figs. 1 and 2). In addition, there are bound to be words that children utter for the first time in multi-word utterances even though they could have been produced in single-word utterances. The percentage of reconstruction, however, was still quite high, as was the case for results obtained using a similar methodology with Hungarian children [12]. The results were clearly better than the simple reproduction of previous children's utterances without allowing for any creative combination, as results go for example

from 68 to 95 % of utterance coverage for the largest corpus tested above (see analysis 2). The improvement in the results was tested in two ways: by enlarging the corpus and by generating new forms around noun and verb categories. Enlarging the corpus provided the best improvement. Enlarging the size of the corpus by a factor of 12 raised the utterance coverage from 87 to 95 % (see analysis 1), but only to 93 % when the noun and verb categories were taken into account. Another important consequence of enlarging the corpus was a much better correspondence between the lexical and syntactic material used by children and their input. Nearly all lexical forms were attested in adult speech (97 %) and a very large part of the syntactic material (80 %) was present and available to be considered as frozen forms. Tomasello [25] emphasizes the importance of having a large and dense corpus to achieve efficient corpus based analyses. For example, adding all previous recordings did not improve the percentages of reconstruction above a certain limit. This could mean that there is a limit to the size of what it is necessary and useful to memorize. It is unclear how far it is possible to go just by increasing the size of the corpus, even using a more varied corpus, but it is unlikely to cover all of children's syntax. Children are capable of producing some type of generalization, and the older they are, the more likely they are to do so. However, the amount of generalization that young children have to use is probably quite low. The linguistic principles proposed here cannot account for all of children's linguistic knowledge. They would produce many aberrant utterances if they were not regulated by other mechanisms. The first of these regulatory mechanisms is semantics, as children produce language that, for them, makes sense. They articulate thoughts with two or three elements that complement each other logically and thus create utterances interpretable by adults. Strange utterances may be produced on occasion but none will sound alien. Secondly, even though children sometimes join words or groups of words randomly when very young, they soon start to follow a systematic order probably copied from adults' utterances [21]. To do this, they merely have to concentrate on the words or groups of words that they already master, having previously uttered them as single words. Indeed, form-function mapping is easier with single-word utterances than with multi-word utterances and this helps to manipulate single-word forms consciously. Thus, single-word utterances are better candidates than most to become the first elements in a combinatorial system and to undergo representational redescription [9]. Their semantic values allow one to perform semantic combinations. By the age of two, associations of words or frozen forms may be sufficient to allow children to produce and control language. The fact that children can learn to produce complex speech patterns quickly without complex grammatical knowledge casts a whole new light on the problem of the acquisition of syntax. The testing procedure relies heavily on semantics because it is assumed that what children understand, they will remember and manipulate. This is perfectly in keeping with recent proposals such as constructivist proposals by Tomasello [24] and Goldberg [6]. More importantly, the possibility of language production by young children without syntactic knowledge changes the fundamental issue of language acquisition. One classical view is that a 'bootstrapping' system is necessary: children need to have some core knowledge of syntax before they can learn the syntax of their mother tongue. Here, there is

no such need as children are able to produce language before actually learning the syntactic regularities and characteristics of the language they hear and use. Another important principle introduced by the construction grammar approach is that adults use a lot of frozen forms and fixed constructions. This means that the mechanism proposed here could remain active with adults, but would become mixed with more complex and sophisticated knowledge as people acquired their mother tongue. It has often been said that children already master syntax by the age of three, which is quite remarkable considering the complexity of what they are acquiring. This report suggests that some simple generative mechanisms can explain the explosive acquisition and apparent mastery of language observed in young children. It demonstrates once again that, as already shown for other linguistic developmental features [5], an apparently complex output may be the product of a simple system. The need for large-scale corpora to better tackle the problem of language acquisition with improved tools is also highlighted here.

Appendix

Table 2 List of missing grammatical elements in the Manchester corpus

3,543	0is	Verb to be, contractions included
1,384	0am	
1,351	0have	
784	0are	
643	0has	
491	0's	Possessive
437	0es	Verb third person singular
160	0s	Plural
121	0ing	
112	0ed	
47	0do	
42	0does	
33	0was	
21	0had	
13	0to	
12	0what	
10	0did	
8	0were	
6	0it	
6	0a	
5	0where	
4	0of	
3	0the	
2	0put	
2	0on	
2	0in	
2	0for	
2	0would	
1	0will	
1	0us	
1	0know	
1	0get	
1	0as	
1	0and	
1	0I	
9,253	Total	

Table 3 Examples of words used in any order within the same recording (children's age ranging from 1;10 to 2;2)

Anne	01B	baby stuck	John	03B	do sock-s
Anne	01B	stuck baby	John	03B	want sock-s do
John	01A	bang bang snail	Liz	03A	that mine that
John	01A	snail bang bang bang			
John	01B	go swim-ing	Anne	04A	fit there down here
John	01B	swim-ing go	Anne	04A	no that fit down there
Warren	01A	Controller gone	Aran	04A	pipe got burst
Warren	01A	gone Controller	Aran	04A	pipe got wet
Warren	01A	there brick there	Aran	04A	look got pipe burst
			Aran	04A	a man there
Aran	02A	Daddy truck	Aran	04A	there a man
Aran	02A	truck Daddy	Aran	04A	me sit there
Aran	02B	toy oh toy there	Aran	04A	sit there me
			Aran	04A	and me sit there
Carl	02A	birdie there no	Aran	04A	it put sand
Carl	02A	no there sheep	Aran	04A	put it that
Dominic	02b	gone train	Carl	04B	car fish
Dominic	02b	train gone	Carl	04B	fish car
			Carl	04B	it dog it eat
Joel	02A	no Mummy			
Joel	02A	no Mummy no	Warren	04A	that one there
			Warren	04A	there that one
John	02B	this it			
John	02B	do it this dolly	Aran	05A	like that
			Aran	05A	that like that
Ruth	02B	baba in there	Aran	05A	that one
Ruth	02B	in there baba	Aran	05A	Daddy get another one that door
			Aran	05A	get get Daddy
Warren	02A	there red there			
Warren	02B	broken it	Carl	05A	Percy no
Warren	02B	it broken	Carl	05A	no Percy
Warren	02B	Warren broken it	Carl	05A	six seven six
Carl	03A	elephant on Thomas	Nic	05A	Mummy no
Carl	03A	there cow on elephant	Nic	05A	no Mummy
Carl	03A	elephant on train			
Carl	03A	hat on man	Ruth	05A	baba eye
Carl	03A	man on horse	Ruth	05A	eye baba
Carl	03A	man on train	Ruth	05B	baba on there
Carl	03A	man on a pink one	Ruth	05B	on there baba
Carl	03A	man on a train	Ruth	05B	Mama baba on there
Carl	03A	man in there man			
Carl	03B	ooh whee	Warren	05A	Mummy look
Carl	03B	whee ooh	Warren	05A	look Mummy
			Warren	05A	a sleep Mummy
			Warren	05A	Mummy sleep Mummy

References

1. Bloom, P. (1999). Theories of word learning: Rationalist alternatives to associationism. In W. C. Ritchie, & T. K. Bhatia, (Eds.), *Handbook of language acquisition*. San Diego: Academic.
2. Brown, R. W. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
3. Chomsky, N. (1959). A review of verbal behavior, by B. F. Skinner. *Language, 35*, 26–58.
4. Clark, E. V. (1993). *The lexicon in acquisition*. New York: Cambridge University Press.
5. Elman, J. L., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press/Bradford Books.
6. Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
7. Ingram, D. (1989). *First language acquisition : Method, description, and explanation*. Cambridge: Cambridge University Press.
8. Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
9. Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press/Bradford Books.
10. Konopczynski, G. (1998). De l'énoncé présyntaxique à la phrase canonique: Aspects syntactico-prosodiques. *Revue PArole, 7–8*, 263–287.
11. Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*, 187–219.
12. MacWhinney, B. (1975). Rules, rote, and analogy in morphological formations by hungarian children. *Journal of Child Language, 2*, 65–77.
13. MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Hillsdale: Lawrence Erlbaum.
14. Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. 2). New York: Gardner Press.
15. Peters, A. M. (1983). *The units of language acquisition*. New York: Cambridge University Press.
16. Peters, A. M. (1995). Strategies in the acquisition of syntax. In P. Fletcher, & B. MacWhinney (Eds.), *The handbook of child language*. Oxford: Blackwell.
17. Pine, J. M. & Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123–138.
18. Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
19. Pinker, S. (1994). *The language instinct: How the mind creates language*. New York/London: William Morrow & Co (New York)/Penguin (London).
20. Ritchie, W. C., & Bhatia, T. K. (1999). Child language acquisition: Introduction, foundations, and overview. In W. C. Ritchie, & T. K. Bhatia (Eds.), *Handbook of language acquisition*. San Diego: Academic.
21. Sinclair, H., & Bronckart, J. P. (1972). S.V.O. a linguistic universal? A study in developmental psycholinguistics. *Journal of Experimental Psychology, 14*(3), 329–348.
22. Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (1999). The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.), *New directions in language development and disorders* (pp. 119–129). New York: Plenum Press. P. 247.
23. Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209–253. P. 8.

24. Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisitiond*. Cambridge, MA: Harvard University Press.
25. Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal Of Child Language, 31*(1), 101–121.
26. Wexler, K. (1982). A principle theory for language acquisition. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition – the state of the art* (pp. 288–315). New York: Cambridge University Press.

Part IV
Linking Syntax to Semantics

Learning to Interpret Novel Noun-Noun Compounds: Evidence from Category Learning Experiments

Barry J. Devereux* and Fintan J. Costello

Abstract The ability to correctly learn to interpret and produce novel noun-noun compounds such as *wind farm* or *carbon tax* is an important part of the acquisition of language in various domains of discourse. One approach to the interpretation of noun-noun compounds assumes that people make use of distributional information about the linguistic behaviour of words and how they tend to combine in noun-noun phrases; another assumes that people activate and integrate information about the two constituent concepts' features to produce interpretations. We present a series of experiments that examine how people acquire both the distributional information and conceptual information that is relevant to compound interpretation. We propose that the relations used to link the two words in noun-noun compounds have rich semantic structure, which includes information about what features of concepts are necessary and/or characteristic for particular relations, as well as distributional information about the frequency with which relations co-occur with different concepts. We present an exemplar-based model of the semantics of relations which captures both of these aspects of relation meaning, and show how it can predict experimental participants' interpretations of novel noun-noun compounds.

*This research was conducted while the first author was a graduate student at University College Dublin.

B.J. Devereux (✉)
Centre for Speech, Language and the Brain, Department of Psychology, University of Cambridge, Cambridge, UK
e-mail: bjd33@cam.ac.uk

F.J. Costello
School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
e-mail: fintan.costello@ucd.ie

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_8,

1 Introduction

People frequently encounter noun-noun compounds such as MEMORY STICK and AUCTION POLITICS in everyday discourse. Compounds are particularly interesting from a language-acquisition perspective: children as young as two can comprehend and produce noun-noun compounds [1], and these compounds play an important role in adult acquisition of the new language and terminology associated with particular domains of discourse. Indeed, most new terms entering the English language are combinations of existing words [2]; consider FLASH MOB, DESIGNER BABY, SPEED DATING and CARBON FOOTPRINT. Noun-noun compounds are also interesting from a computational perspective, in that they pose a significant challenge for current computational accounts of language understanding. This challenge arises from the fact that the semantics of noun-noun compounds are extremely diverse, with compounds utilizing many different linking relations between their constituent words [3–5]. Despite this diversity, people typically interpret even completely novel compounds extremely quickly.

One approach that has been taken in both cognitive psychology and computational linguistics can be termed the *relation-based approach* (e.g. [6, 7]). In this approach, the interpretation of a compound is represented as the instantiation of a relational link between the modifier and head noun of the compound. Such relations are usually represented as a taxonomy; for example the meaning of STUDENT LOAN might be specified with a POSSESSOR relation [7] or MILK COW might be specified by a MAKES relation [6]. However, researchers are not close to any agreement on a taxonomy of relation categories classifying noun-noun compounds; indeed a wide range of typologies have been proposed (e.g. [4, 7]).

In these relation-based approaches, extrinsic linguistic information about the concept terms, such as distributional information about how often different relations are associated with a concept word, is taken to be the influential factor influencing the interpretation process, and there is often little focus on how the meaning of the relation interacts with the intrinsic properties of the constituent concepts (on the distinction between intrinsic and extrinsic features, see [8, 9]). For example, the CARIN model [6] utilizes the fact that the modifier MOUNTAIN is frequently associated with the LOCATED relation (in compounds such as MOUNTAIN CABIN or MOUNTAIN GOAT); the model does not utilize the fact that the concept MOUNTAIN has intrinsic properties such as *is large* and *is a geological feature*: features which may in general precipitate the use of a LOCATION relation.

An approach that is more typical of psychological theories of compound comprehension can be termed the *concept-based approach* [10, 11]. For these accounts, the focus is on the intrinsic properties of the constituent concepts, and the interpretation of a compound is usually modelled as a modification of the head noun concept. So, for example, the compound ZEBRA FISH may involve a modification of the FISH concept, by asserting a feature of the ZEBRA concept (e.g. *has stripes*) for it; in this way, a ZEBRA FISH can be understood as a fish with stripes. Concept-based theories

do not typically use distributional information about how various relations are likely to be used with concepts.

Thus, the information assumed relevant to compound interpretation is quite different in relation-based and concept-based theories. However, neither approach typically deals with the issue of how people acquire the information that allows them to interpret compounds. In the case of the relation-based approaches, for example, how do people acquire the knowledge that the modifier MOUNTAIN tends to be used frequently with the LOCATED relation, and that this information is important in comprehending compounds with that modifier? In the case of concept-based approaches, how do people acquire the knowledge that particular features of ZEBRA are likely to influence the interpretation of ZEBRA FISH?

We present experiments which examine how both distributional information about relations and intrinsic information about concept features influence compound interpretation. We also address the question of how such information is acquired: our experiments used laboratory-generated concepts that participants learn during the experiments. As well as learning novel concepts, participants also learn how these concepts tend to combine with other concepts via relational links. Using laboratory-controlled concepts allows us to control and manipulate various factors that might be expected to influence compound comprehension; for example, concepts can be designed to vary in their degree of similarity to one another, to be associated with potential relations with a certain degree of frequency, or to have a feature which is associated with a particular relation. It would be extremely difficult to control for such factors, or investigate the acquisition process, using natural, real world concepts. Using laboratory-generated categories also eliminates confounding factors such as lemma and compound frequency which have often proved contentious in studies of natural language compounds [12, 13].

2 An Exemplar-Based Account of Compound Interpretation

As mentioned above, one characteristic of the relation-based approach is the view that the interpretation of compounds can be represented using a taxonomy of relation types. Models typically assume that the semantics of the relations used in compound interpretation can be adequately represented using a set of relation labels (e.g. *located*, *for*) for which no internal semantic structure is posited (other than their association with the modifier term, e.g. [6]). In this chapter we take a different view, and propose that relations are at least as complex as the concepts which they link in terms of the representational demands that they put on semantic memory. Consistent with much of the work on the representation and meaning of concepts (e.g. [14–18]), we will assume that relations vary in terms of their semantic richness, their internal complexity, and in their similarity to each other. Our key claim therefore is that the relations used in compounds are complex representations and that a successful model of compound interpretation must account for how the appropriate relation representation becomes activated during the interpretation process.

The exemplar theory ([19–22]) of categorization is one model that assumes a rich representational framework for conceptual knowledge: each concept is represented as a set of individual memory traces (exemplars) which in turn are represented as attributes on a fixed set of dimensions. In our model, we assume that relations can be represented using an exemplar approach in the same way concepts are in the exemplar theory. Our framework for representing relations is therefore not based on a taxonomy of relation labels but instead uses representations as rich as those typically proposed for conceptual knowledge.

Our key theoretical claim is that the interpretation of a compound consists of activating exemplars of the two constituent concepts in the compound, which in turn activates the relational exemplars with which they are associated. Just as each exemplar in a conceptual category consists of a list of attributes, for us each exemplar in a relational category shall consist of the attributes of the two exemplars of the conceptual categories for which the relation is instantiated. Each relation exemplar is a memory trace unique to the situation in which it is instantiated. For example, if A and B are two conceptual categories consisting of the set of exemplars $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ and $\{b_1, b_2, b_3, b_4, b_5\}$ respectively, then a relation R that can be used to link the concepts A and B can be represented as the set of instances for which that relation holds (e.g. $\{(a_1, b_1), (a_2, b_2), (a_3, b_3), (a_4, b_5), (a_5, b_5)\}$). Important aspects of the relation semantics are captured by the features of the concept exemplar pairs for which the relation holds. To take a real-world example, if an *is standing upright on* relation occurs between a KETTLE exemplar and a SHELF exemplar, a particular aspect of the semantics of this relation is captured by the fact that a feature of one exemplar (i.e. the flat base of the kettle) is in physical contact with a feature of the second exemplar (i.e. the flat horizontal surface of the shelf).

In this exemplar-based approach, we can compute the appropriateness of the relation R for two arbitrary conceptual exemplars x and y by computing the membership of (x, y) in category R, using a standard exemplar modelling framework implementing the exemplar theory of category learning. If more than one relation is defined, we can compute membership in each of the relations and thus make predictions about how likely each relation is for the pair of items.

3 Overview of Experiments

Assuming a rich framework for representing the meaning of relations and their associations with the features of concepts allows us to investigate important issues in compound interpretation that have not typically been investigated in the literature. In the first two experiments, we investigate the notion that the presence of certain semantic features in the representation of a constituent concept can be necessary for the instantiation of a particular relational link. For example, the compound CUPBOARD FRUIT can be understood as *the fruit located in the cupboard* by instantiating an H *is located inside* M type of relation between the modifier (M) and head (H) of the phrase, but such an instantiation requires that certain features

must be present; in particular, the modifier concept must be a kind of container. A *during* relation cannot be instantiated for CUPBOARD FRUIT because the property required to facilitate that relation (namely, the property of being a period of time) is not present in the modifier. Thus, the instantiation of some relational links between concepts is impossible if certain necessary properties (which we refer to as *facilitating features*) are absent from either the modifier concept or the head-noun concept. Facilitating features are part of the intensional representation for the relation and are independent of the relation's extensional representation (i.e. the distribution of head and modifier terms for instances of the relational link) which is of primary importance in relation-based approaches such as the CARIN model. Facilitating features represent hard constraints on relation selection.

The notion of facilitating features in Experiments 1 and 2 has much in common with the componential view of verb meaning that has be proposed in the psycholinguistic literature, where a verb like GIVE is represented in terms of components such as *cause*, *do*, *change*, *possession* which hold between an object, a giving agent and a receiving agent [23]. A prerequisite (i.e. facilitating feature) for GIVE is that the agent must possess the object to be given [23]. Furthermore, facilitating features are similar to semantic selectional restrictions on verb subcategorization frames, which impose constraints on what kinds of objects can populate given argument slots for particular verbs (e.g. for the verb EAT, the object in the direct object slot must be edible). As such, facilitating features in compound interpretation can be seen as falling into a more general framework implicating semantic restrictions in relational processing in language.

Related ideas have also been considered in the domain of problem-solving. For example, in the radiation problem [24], *destructive force* is identified as a *functionally relevant attribute* of the rays which can attack the tumor, which must also be true in the analogous scenario if the correct analogical mapping is to be found [25]. Functionally relevant attributes correspond to the facilitating features in the constituent concepts of noun-noun compounds, as in both domains the features are necessary for the selection of the appropriate relation. Experiments 1 and 2 test the idea that relations often require the presence of facilitating features, and that which relation is used to interpret a compound is influenced by the presence or absence of facilitating features in the constituent concepts.

Functionally relevant attributes may or may not be salient for a concept [25]. In the same way, we can differentiate between facilitating and salient features. We use *diagnosticity* as an operational measure of feature saliency; the diagnosticity of a feature for a category is a measure of how useful that feature is for identifying members of the category; a feature is highly diagnostic of a category if it occurs in many instances of that category and in few instances of other categories [26]. Facilitating features are always at least somewhat diagnostic of their associated relation, because they necessarily occur in every instance of that relation. However, facilitating features can also be relatively undiagnostic of a particular relation if they also occur in many instances of other relations. Unlike facilitation, diagnosticity represents a soft constraint on what relations are possible given the features of the constituent concepts, representing as it does the statistical dependencies between

relations and features. Though diagnosticity and facilitation tend to overlap, we manipulate them independently in our experiments.

The theory outlined in Sect. 2 proposes that the interpretation of novel compounds is derived from activation of the relational exemplars that are associated with the exemplars of the two constituent concepts in the compound. This proposal was described in nonlinguistic terms: interpretation is based on conceptual and relational semantic knowledge associated with the two constituent concepts. In particular, we did not propose a role for syntax (i.e. the order of the two words in the compound) in determining which relation exemplars become activated. For example, our account predicts that the same relation is instantiated for the novel compounds FRUIT CONTAINER and CONTAINER FRUIT – in both cases the link between the concept CONTAINER and the concept FRUIT is that the fruit is in the container – although the head of the two phrases differ (i.e. a fruit container is a type of container and container fruit is a type of fruit).[1] Other theories (e.g. the CARIN theory) propose that the modifier concept plays a crucial role in determining how relations are selected during the interpretation process. Experiments 2 and 3 aim to investigate the role of syntax on the selection of relations for compounds by manipulating the order in which the two terms in the compound appear. Although our theory and models do not predict differences due to word order, such differences would indicate that participants are applying syntactic processing to the compound stimuli and are not combining the learned category names in non-linguistic ways.

For each of our experiments, we present an exemplar-based model of the relation selection process. To foreshadow our results, our models give a good fit to data on how compounds featuring novel concepts that are learned during the experiment are interpreted by participants.

4 Experiment 1

Experiment 1 aims to test the hypothesis that relations are meaningfully represented by an exemplar category structure. This is a separate issue to the one of how an exemplar representation of relations might account for noun-noun compound comprehension, and Experiment 1 does not examine compound interpretation directly. Experiment 1 focuses on whether people can learn different relations through experience of the relations linking pairs of items, and if so can they use their learning to make judgements about which relations are plausible or likely to hold between other pairs of items. In the experiment, artificial laboratory-generated relation categories are constructed and experimental participants must learn how

[1] There are exceptions where the type of relation differs depending on word order; these tend to be lexicalized compounds, or compounds containing a polysemous word where the sense in the modifier position can differ from the sense in the head position (e.g. GUITAR SOLO and SOLO GUITAR).

to distinguish between them. Our experiment therefore has much in common with experiments presented in the category learning and classification literature (e.g. [19, 22]; see [27] for an overview): a preliminary training phase where participants are exposed to exemplars of different categories is followed by a transfer phase where participants judge the category membership of new items.

The categories are four different relations that can hold between pairs of objects and each of the training items consist of two objects linked by one of these relations. When learning about the relation categories, information that participants are required to attend to includes information about the facilitating features of relations (some features are facilitating for one of the four relations; some are not prerequisites for any relation). The experiment thus addresses the question of whether people are sensitive to the differences between facilitating and non-facilitating features.

Features also vary in whether they are diagnostic for a given relation. The *diagnosticity*, D, of a feature f for a category C can be defined as:

$$D_C(f) = \frac{|E_C \cap E_f|}{|E_C \cup E_f|} \tag{1}$$

where E_f denotes the set of exemplars that have feature f and E_C denotes the set of exemplars that belong to category C [11]. A feature has maximal diagnosticity for a relation if that feature is present in every instance depicting that relation and not present in every instance that does not depict that relation, and diagnostic features are very characteristic of their associated relation. Facilitating features necessarily have some degree of diagnosticity for their corresponding relations because they must appear in every instance of that relation (they may also appear in some instances of other relations). Non-facilitating features can be either diagnostic or non-diagnostic for relations. We would expect participants to attend to these diagnostic features when learning the training items and make use of them when selecting relations for the new transfer items, consistent with findings in the categorization literature [28–30].

To summarize, the aim of the experiment is to assess whether people can learn relations from a set of training items, how they would use their learning to rate new items, and whether they are especially influenced by features that are facilitating for relations – would participants employ different strategies with respect to facilitating and non-facilitating features?

4.1 Method

4.1.1 Participants

Sixteen postgraduate students or recent college graduates volunteered to take part in the experiment. All were native speakers of English.

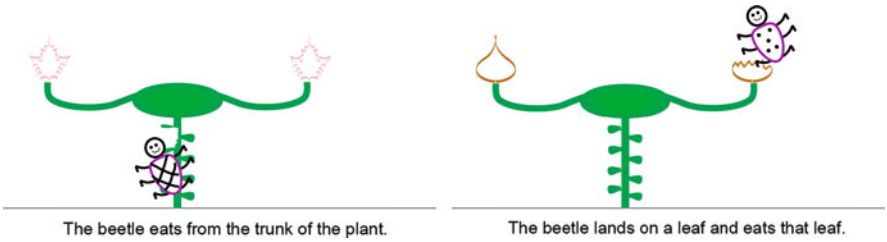

Fig. 1 Two examples of stimuli used in the training phase of Experiment 1

4.1.2 Materials

The training items consisted of 18 visual stimuli depicting an imaginary beetle eating an imaginary plant, with each stimulus presented on an A5-sized card (Fig. 1; similar depictions have been used in other category learning studies [31, 32]). The beetles varied on three feature dimensions: colour of shell, pattern on shell, and facial expression. The plants varied on four feature dimensions: colour of leaves, shape of leaves, droop of branch, and whether there were buds or thorns on the trunk. There were four possible ways in which a beetle could eat a plant: the beetle could land on a leaf of a plant and eat the leaf, the beetle could eat from the top of the trunk of the plant, the beetle could eat from the trunk of the plant if there were buds rather than thorns on the trunk, or the beetle could stand on the ground and eat the leaf of a plant that had drooping branches. These four different types of eating behavior were the concrete manifestation of the four relation categories used in the experiment. Underneath each training picture was a sentence describing the eating behavior talking place.

The 18 training items had the abstract category structure shown in Table 1. The four relations are partitioned into two types: Relations 1 and 2 are what we term *independent relations*: they are possible regardless of what features are present in either the plant or the beetle. Relations 3 and 4 are *dependent relations*: whether these relations are possible is contingent on the presence of certain plant features. The concrete relation "the beetle stands on the ground and eats the leaf" is only possible if the branches of the plant are drooping, while the concrete relation "the beetle eats from the trunk of the plant" is only possible if there are buds rather than thorns on the trunk of the plant.

There are seven feature dimensions; B1, B2 and B3 denote the three abstract beetle dimensions (corresponding to the concrete features of colour, pattern and facial expression). Pf1 and Pf2 denote the two facilitating plant dimensions: they correspond to the features that are required for one of the relations to be possible, namely drooping of branches and buds or thorns on the trunk. Pf1 is the facilitating dimension of Relation 3: this relation is only possible if there is a 2 on Pf1. Pf2 is the facilitating dimension of Relation 4: this relation is only possible if there is a 2 on Pf2. P1 and P2 denote the two abstract plant dimensions that do not facilitate dimensions (corresponding to the concrete features of leaf colour and leaf shape).

Table 1 The abstract relational category structures used in the training phase

		Beetle features			Plant features			
Item	Relation	B1	B2	B3	Pf1	Pf2	P1	P2
1	Relation 1	1	1	1	1	2	1	1
2	Relation 1	4	1	1	1	1	1	1
3	Relation 1	2	2	2	1	2	4	1
4	Relation 1	3	4	2	2	1	1	4
5	Relation 1	3	3	3	1	1	1	1
6	Relation 2	1	1	2	1	2	3	2
7	Relation 2	2	2	2	1	1	3	2
8	Relation 2	2	3	4	1	1	1	1
9	Relation 2	3	4	3	2	2	2	3
10	Relation 3	1	4	1	2	1	4	2
11	Relation 3	1	1	4	2	1	2	2
12	Relation 3	2	2	2	2	2	2	3
13	Relation 3	3	2	4	2	1	2	3
14	Relation 3	2	2	3	2	1	3	2
15	Relation 4	1	3	1	2	2	4	3
16	Relation 4	2	3	3	1	2	2	2
17	Relation 4	3	2	3	1	2	3	3
18	Relation 4	3	3	3	1	2	3	3

Dimensions P1 and P2 are important for identifying items that belong to the Relation 1 category as for each of these dimensions a 1 occurs four out of five times within the category but only one out of 13 times outside the category; i.e. a 1 on P1 and a 1 on P2 are diagnostic features for Relation 1. There is no such diagnostic feature for Relation 2; the best example of a diagnostic feature for this relation is a 3 on P1, which occurs only two out of four times within the category and three out of 14 times outside the category. A 2 on Pf1 is already important for identifying items that belong to the Relation 3 category as it is that relation's facilitating feature. However Pf1 is of added importance because a 2 on Pf1 is also a very diagnostic feature for the relation, occurring five out of five times within that category and only three out of 13 times outside it. In contrast, a 2 on Pf2, the facilitating feature for Relation 4, is not as diagnostic for that relation, occurring four out of four times within the category but five out of 14 times outside it. A 3 on B2, a 3 on B3, and a 3 on P1 are each moderately diagnostic features for Relation 4. A 2 on P1 is also moderately diagnostic for Relation 3.

The actual mappings of the abstract beetle and plant features to the concrete beetle and plant features was varied across participants. For half of the participants Relation 3 was mapped to the concrete relation "the beetle stands on the ground and eats the leaf" while Relation 4 was mapped to "the beetle eats from the trunk of the plant"; for the other half of participants these assignments were reversed. For the participants that saw Relation 3 as the beetle eating from the ground and

Table 2 The abstract features to concrete features mappings for the 16 experimental participants in Experiment 1

Relations 3 and 4	Relations 1 and 2	Plant features	Subj.
Relation 3 is *beetle stands on the ground and eats the leaf*; Relation 4 is *beetle eats from the trunk of the plant*.	Relation 1 is *beetle eats from leaf*; Relation 2 is *beetle eats from oval mound at top of trunk*	P1 is *leaf shape*; P2 is *leaf colour*	1 2
		P1 is *leaf colour*; P2 is *leaf shape*	3 4
Pf1 is *droop/no-droop of branches*; Pf2 is *buds/thorns on trunk*	Relation 1 is *beetle eats from oval mound at top of trunk*; Relation 2 is *beetle eats from leaf*	P1 is *leaf shape*; P2 is *leaf colour*	5 6
		P1 is *leaf colour*; P2 is *leaf shape*	7 8
Relation 4 is *beetle stands on the ground and eats the leaf*; Relation 3 is *beetle eats from the trunk of the plant*.	Relation 1 is *beetle eats from leaf*; Relation 2 is *beetle eats from oval mound at top of trunk*	P1 is *leaf shape*; P2 is *leaf colour*	9 10
		P1 is *leaf colour*; P2 is *leaf shape*	11 12
Pf2 is *droop/no-droop of branches*; Pf1 is *buds/thorns on trunk*	Relation 1 is *beetle eats from oval mound at top of trunk*; Relation 2 is *beetle eats from leaf*	P1 is *leaf shape*; P2 is *leaf colour*	13 14
		P1 is *leaf colour*; P2 is *leaf shape*	15 16

Relation 4 as the beetle eating from the trunk, the facilitating plant dimensions Pf1 and Pf2 were mapped to the concrete features of drooping/non-drooping branches and buds/thorns on trunk respectively, while the other group of participants saw Pf1 mapped to buds/thorns on trunk and Pf2 mapped to drooping/non-drooping branches respectively; this was necessary to ensure that the dependent relations were associated with the correct facilitating features. Within each of these two groups, the assignments of the abstract dimensions P1 and P2 to the concrete dimensions of leaf color and shape were also counterbalanced, and the mappings of the abstract values to the concrete values on both of these dimensions were randomized for every participant. Half of the participants saw Relation 1 mapped to the concrete relation "the beetle lands on a leaf and eats the leaf" and Relation 2 mapped to the concrete relation "the beetle eats the oval mound at the top of the trunk" and the other half of the participants saw the opposite mapping. For the beetle features, mappings were generated for each participant by randomly assigning the three abstract dimensions to the three concrete dimensions and then randomly assigning the four possible abstract values within each dimension to each of the four concrete values for that dimension. Table 2 summarizes these mappings for the 16 participants.

The materials for the transfer phase consisted of more pictures depicting beetles and plants; however in these pictures the beetles and plants were shown separately,

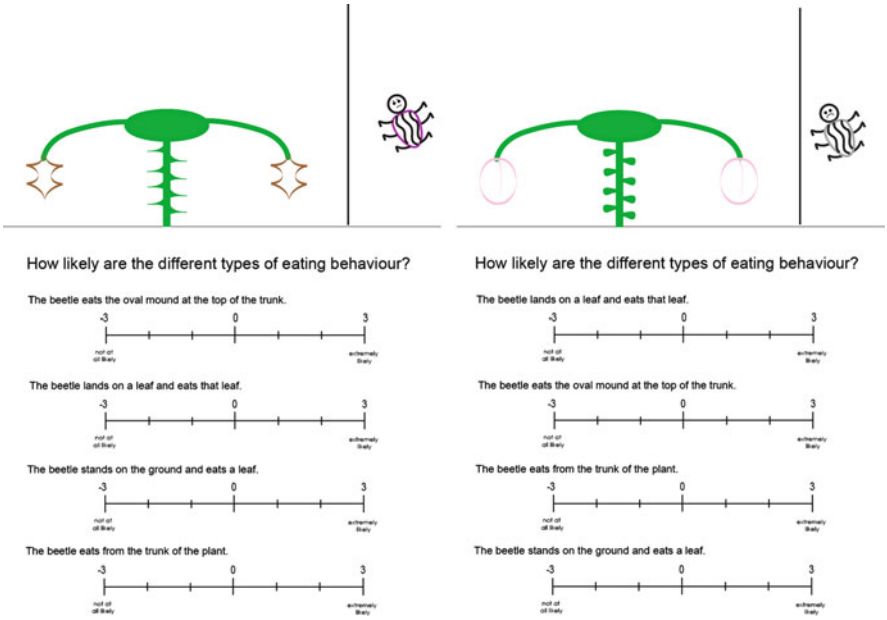

Fig. 2 Two examples of stimuli used in the transfer phase of Experiment 1

without eating relations taking place (Fig. 2). Underneath each picture was the question "How likely are the different types of eating behavior?" followed by four relation description sentences, each of which was accompanied by a seven-point scale. The order of the scales on the pages was counterbalanced across participants.

There were two stages in the transfer phase of the experiment. In the first stage participants were tested with nine of the 18 training items (two items from each of the four relational categories, with the ninth item selected from Relation 1; the items were 1, 4, 5, 7, 9, 12, 13, 15 and 18 in Table 1). These training items reoccurred in the transfer phase of the experiment in order to have an effective method of measuring how accurately participants learned the training items.

Twenty new transfer items were also constructed (Table 3). These items were designed to test what effect the presence or absence of the facilitating features for the Relation 3 and 4 (i.e. values on Pf1 and Pf2) would have on people's relation selections. The items also varied in how easy it would be to select a relation; the selection of a relation should be relatively easy for an item if it is very similar to a training item for that relation (i.e. differing from a training item on only one dimension) and if it is very prototypical of that relation (ignoring values on dimensions Pf1 and Pf2). Half of the 20 transfer items are of this form (items 1, 2, 3, 4, 9, 10, 13, 14, 17 and 18). The other ten transfer items were constructed so that the selection of a relation should be difficult, in the sense that they were dissimilar to the training items (the average number of common features with the most similar training item was 4.0) and tended to manifest conflicting evidence

Table 3 The 20 new exemplar items used in the transfer phase of Experiment 1

	Beetle features			Plant features			
Item	B1	B2	B3	Pf1	Pf2	P1	P2
1	3	1	1	1	1	1	1
2	2	4	2	1	1	3	2
3	3	2	4	1	1	2	3
4	3	3	3	1	1	3	3
5	1	4	4	1	1	4	4
6	4	3	2	1	1	2	4
7	1	2	4	1	1	3	3
8	2	4	2	1	1	1	4
9	1	2	4	2	1	2	2
10	4	1	1	2	1	1	1
11	1	4	4	2	1	4	4
12	3	1	1	2	1	4	4
13	2	3	3	1	2	3	3
14	2	1	2	1	2	3	2
15	1	4	4	1	2	4	4
16	3	3	3	1	2	2	2
17	2	2	4	2	2	2	3
18	3	3	3	2	2	3	3
19	1	4	4	2	2	4	4
20	3	3	3	2	2	4	4

for two different relation categories (for example, in item 16, there is conflicting evidence for Relations 3 and 4), or exhibit little or no evidence for any relational category. The ways in which the 20 new transfer items differed from each other with respect to these factors is summarized by Table 4.

4.1.3 Procedure

As described above, the experimental procedure consisted of three sections: a training phase where participants studied the training items, a test phase where they had to select relations for some of the beetle-plant pairs seen in training, and a transfer phase where they rated new items. In the training stage, participants were instructed to pretend to be biologists interested in learning about imaginary plants and beetles and in particular the kinds of eating behaviour that existed between different plants and beetles. Participants were given the 18 training items in a random order at a large desk area, and were given 12–15 min to study the items. After the training phase, the training items were removed and participants were given the nine test items. Participants were told to indicate the likelihood of the various ways of eating for each beetle-plant pair, using the four scales provided. After participants had rated these items the 20 transfer items were presented. In

Table 4 Summary of the construction of the 20 transfer items for Experiment 1

Pf1	*Pf2*	Classification ease	Item
Pf1 is not facilitating	*Pf2* is not facilitating	Easy to classify (similar to a training item, prototypical of a relation)	1
			2
			3
			4
		Difficult to classify (dissimilar, little evidence for any relation or conflicting evidence for two or more relations)	5
			6
			7
			8
Pf1 is facilitating	*Pf2* is not facilitating	Easy to classify (similar to a training item, prototypical of a relation)	9
			10
		Difficult to classify (dissimilar, little evidence for any relation or conflicting evidence for two or more relations)	11
			12
Pf1 is not facilitating	*Pf2* is facilitating	Easy to classify (similar to a training item, prototypical of a relation)	13
			14
		Difficult to classify (dissimilar, little evidence for any relation or conflicting evidence for two or more relations)	15
			16
Pf1 is facilitating	*Pf2* is facilitating	Easy to classify (similar to a training item, prototypical of a relation)	17
			18
		Difficult to classify (dissimilar, little evidence for any relation or conflicting evidence for two or more relations)	19
			20

both the test and transfer stages, the order in which the items were presented was randomized for each participant, and participants were allowed rate the items at their own pace.

4.2 Results

4.2.1 Performance on the Nine Test Items

For the nine test items there was a "correct" answer; that is, the relation they had appeared with in the training stage. Analysis of these nine test items allows us to quantify how well participants learned the training items. The responses for

each relation and each experimental item were classified as either positive (>0) or non-positive (≤0), depending on how the participant responded on each scale. On average, participants gave a positive rating to the correct relation for 70.8 % of the test items (recall). The proportion of positive responses that were to correct relations was 37.9 % (precision). For the incorrect relations, only 38.7 % of the responses were positive. These results indicate that participants learned to distinguish between the categories in the training phase.

4.2.2 Sensitivity to Facilitating Features

In total, participants responded to 29 items; of these, 16 were items for which the facilitating feature for Relation 3 was absent and 16 were items for which the facilitating feature for Relation 4 was absent. For the 16 items for which the facilitating feature for Relation 3 was absent, each participant made a response for Relation 3. For each participant, the proportion of these responses that were non-positive was the statistic of interest. A one-tailed binomial test was calculated for each participant. The test was significant for 14 participants; in other words, 14 of the 16 participants were significantly more likely to produce a non-positive rather than a positive response to Relation 3 when the facilitating feature for Relation 3 was absent. (Indeed, 11 participants never produced a positive response). For Relation 4, 11 of the 16 participants were significantly more likely to produce a non-positive rather than a positive response. This is strong evidence that participants do not consider the dependent relations to be possible for items in which the facilitating feature is absent.

A similar analysis was performed looking at the items where the facilitating feature was present. Of the 29 items, 13 were items for which the facilitating feature for Relation 3 was present and 13 were items for which the facilitating feature for Relation 4 was present. For Relation 3, 9 of the 16 participants were significantly more likely to produce a positive rather than a non-positive response. For Relation 4, 7 of the 16 participants were significantly more likely to produce a positive rather than a non-positive response. Therefore, many participants often rate a relation as having low likelihood for a given item, even when that relation's facilitating feature is present in the item. This makes sense: the presence of a facilitating feature does not entail that the relation must be selected for the item, only that it is a possibility, and participants may notice greater evidence for a different relation.

The responses for Relations 3 and 4 for items with facilitating features present and facilitating features absent were also analysed by means of a 2×2 ANOVA. In the by-subject analysis, Relation (3 or 4) and facilitation (present or absent) were within-subject factors. In the by-item analysis, Relation (3 or 4) was a within-item factor and facilitation (present or absent) was a between-item factor. There was a main effect of facilitation ($F_s(1, 15) = 62.58,\ MSE = 2.11,\ p < 0.001$; $F_i(1, 28) = 127.07, MSE = 1.17, p < 0.001$) indicating that participants were very attentive to the facilitating nature of the relevant feature. There was no main effect of Relation ($F_s,\ F_i < 1$) or interaction between facilitation and response relation.

Overall, the results indicate that the participants were extremely sensitive to the effects of the facilitating properties.

4.2.3 Sensitivity to Diagnostic Features

Using Eq. 1, we calculated the average diagnosticity of the features of each of the 29 test items for each of the four relational categories and compared this to the observed data. For two relations, the average diagnosticity for items had a high correlation with the observed membership ratings for the items (for Relation 1, $r = 0.83$, $p < 0.01$; for Relation 4, $r = 0.81$, $p < 0.01$). For the other two relations, the correlation was less strong though still significant (for Relation 2, $r = 0.66$, $p < 0.01$; for Relation 3, $r = 0.70$, $p < 0.01$). These results indicate that participants were sensitive to the diagnosticity of features as well as whether or not they were facilitatory when making their category judgments.

4.2.4 Modelling Relation Selection

Our account of the process by which noun-noun compounds are interpreted (Sect. 2) posits the activation of a relational link between the two concepts, which proceeds from the retrieval of the relational exemplars that tend to co-occur with the exemplars of the two constituent concepts in the compound. These retrieved relational exemplars form the basis for interpretation. This account predicts that participants will base their judgments for novel exemplar pairs on how similar exemplars were classified in the training phase of the experiment. This is consistent with how classification is assumed to proceed in exemplar models of categorization such as the Generalized Context Model (GCM) [20, 21]. Our proposal here is that finding a plausible relation for a pair of items can be modelled as an exemplar categorization process.

We modelled the data using the GCM.[2] This model computes the probability of item i belonging in category C as a choice function of exemplar similarity:

$$P(i, C) = \frac{\sum_{j \in C} sim(i, j)}{\sum_{j \in U} sim(i, j)} \tag{2}$$

[2]Other exemplar modelling frameworks, such as the Diagnostic Evidence Model [28] and TiMBL [33] could also have be investigated. However, comparing different modelling frameworks on this task lies beyond the scope of this chapter.

where U denotes the set of all exemplars and $sim(i, j)$ is the measure of similarity between exemplars i and j. Similarity between exemplars is in turn defined as negative-exponential transformation of distance [34]:

$$sim(i, j) = e^{-c \times dist(i,j)} \tag{3}$$

where c is a free parameter specifying how quickly similarity between exemplars diminishes as a function of distance. The distance between two exemplars is usually measured by a metric such as the L_2 (Euclidean) norm or the L_1 norm [21, 35, 36]; however such metrics assume dimensions ranging over real-valued intervals for which a modulus is defined. Since the values of our dimensions (e.g. different leaf shapes) are not interval-valued we define distance as

$$dist(i, j) = \sum_{d \in D} w_d m_{dij} \tag{4}$$

where D is the set of seven dimensions and m_{dij} is a discrimination function, evaluating to 0 if exemplars i and j have the same value on dimension d and 1 otherwise. The w_d are a set of attentional weights satisfying the constraints $0 \leq w_d \leq 1$ and $\sum w_d = 1$. $dist(i, j)$ defines a metric (a Hamming distance with weights on the string positions) and is a generalization of the L_1 metric on binary-valued dimensions, often used in the classification literature (e.g. [27]).

The data collected in Experiment 1 correspond to individual subjects' belief in the likelihood of each relation for each item (i.e. a Bayesian measure). However, the GCM models frequency probability, i.e. the probability that a participant will select a particular relation as being correct for a given item. To create a probabilistic measure from our data which can be modelled with the GCM, we assumed that when a participant rates a particular relation as being more likely than the others, this is the relation that the participant would select in a forced choice paradigm.[3] The GCM utilises seven free parameters (the scaling parameter c and the seven dimensional weights with six degrees of freedom); the parameter values that maximised the correlation, across all items and relations, to the empirical data were estimated using a brute-force search over the parameter space. The GCM provides a reasonable fit to the data (for Relation 1, $r = 0.89$, $p < 0.01$; for Relation 2, $r = 0.69$, $p < 0.01$; for Relation 3, $r = 0.90$, $p < 0.01$; for Relation 4, $r = 0.94$, $p < 0.01$).

However, in this form, the GCM does not distinguish between facilitating and non-facilitating features. Our results show that participants use their knowledge of the world when deciding whether or not the facilitating properties prohibit a particular relation. We model this by assuming that when participants dismiss a particular relation because of the absence of a facilitating feature, they make an *a priori* judgment that is independent of any later exemplar-driven process and

[3] If a participant rates two or more relations with the same maximal likelihood, we assume the participant would select at random between them. Whether or not the data are actually transformed in this way makes only small differences to the fit of the model reported subsequently.

therefore ignore those training items they remember which depicted a relation that is impossible for the transfer item at hand. We modified the GCM to account for the effect of the facilitating relations so that membership of an item in a relational category is not computed from the complete set of learned exemplars, but rather from the subset of the learned exemplars that do not belong to relational categories that are impossible for the current item. The choice function then becomes

$$P(i, C) = \frac{\sum_{j \in C \cap A_i} sim(i, j)}{\sum_{j \in U \cap A_i} sim(i, j)} \tag{5}$$

where A_i denotes the set of exemplars in memory which do not belong to relations incompatible with the facilitating properties of item i. In this form, the GCM gives items zero probability for membership in relational categories that are impossible because of the absence of facilitating properties. This model does indeed give a closer fit to the data (for Relation 1, $r = 0.90$, $p < 0.01$; for Relation 2, $r = 0.89$, $p < 0.01$; for Relation 3, $r = 0.92$, $p < 0.01$; for Relation 4, $r = 0.94$, $p < 0.01$).

4.3 Discussion

Experiment 1 showed that people can learn which relations hold between concepts from sets of examples of those relations and pay attention to both facilitating features and the diagnosticity of features for relations when judging relation likelihood for new examples. That an exemplar model of classification accurately models how people rate the likelihood of relations holding between pairs of objects is evidence in support of the hypothesis that relations can be modelled as exemplars, much in the same way conceptual categories can be. Such findings are consistent with our claim that the selection of a relation for two constituents can be seen as a kind of exemplar categorization task, with decisions about category membership computed from the access of relevant memory traces in memory.

These findings have implications for theories on the role of relational links in conceptual combination, such as the CARIN model [6]. Since the thematic relations used in the CARIN model have no internal structure there is no way in which facilitating or diagnostic features could be part of the representation of relations in that model (for example, the *H made of M* relation in the CARIN model has no way of requiring that a concept taking part in it is type of substance). Concept modification approaches do typically allow for internal conceptual structure to influence relation selection [10,11]. However, the exemplar-based model of relation selection described here provides an alternative to the slot-based representation of relations typically assumed in concept modification models, showing that relations

can instead be represented as sets of paired-item exemplars. This exemplar-based model has the advantage of giving a simple account of how people learn which properties are associated with each relation.

However, compound comprehension was not addressed directly in this experiment: participants made judgments of relation likelihood for pairs of exemplars rather for noun-noun compound phrases. Participants learned relational categories only, and did not learn conceptual categories that could be used as the words in the compounds (i.e. they rated relations for exemplar pairs e_i and e_j rather than for noun-noun compounds of the form "A B", where A and B are the names of concepts). In Experiments 2 and 3, we aim to extend the methodology of Experiment 1 to noun-noun compound interpretation. The goal of these experiments is to reveal more both about how relations are learned and used and about how exemplar-level and conceptual-level information interact during conceptual combination.

5 Experiment 2

In Experiment 2 participants learn different conceptual categories (i.e. different types of beetles and plants) as well as different relational categories. In the transfer phase, participants are presented with a pair of beetle and plant category labels (in the form of a noun-noun compound) and are required to make judgments about which of the different relations are likely or appropriate for that compound. Experiment 1 showed that participants can learn about the relations between beetle and plant items and use that information to make judgments about the likelihood of relations for new items; Experiment 2 examines whether that learning can be generalised to a noun-noun comprehension task.

For real-world concepts and relations, features vary in their diagnosticity for both relations and the concepts that they link. The feature *has three wheels*, for example, is diagnostic of TRICYCLE but is not diagnostic of any particular relation: it is difficult to imagine a relational link between two concepts that is dependent on one of the concepts having three wheels specifically (i.e. if the property *has three wheels* is untrue then the relation is impossible). Conversely, the feature *has a horizontal flat surface raised off the ground* is not particularly diagnostic of any concept, though it is diagnostic of the *stands upright on* relation, as the feature is in fact a facilitating feature for that relation. We wished to capture similar qualities in the design of the abstract category structure used in Experiment 2. Sometimes a beetle feature is highly diagnostic of a beetle category and less diagnostic of a relation category; sometimes a beetle feature is highly diagnostic of a relation category and less diagnostic of a beetle category.

The design incorporates a single facilitating feature/facilitated relation pair; beetles could eat from the trunk of a plant only if the trunk was free of thorns. The two plant categories are designed to be identical in terms of how difficult they are to learn and how frequently they occur with the various relations but differ in terms of how often a feature that is facilitating for a particular relation occurs

amongst their exemplars. That is, the facilitating feature occurred more often in one plant category than in the other; the feature varied in its diagnosticity for the plant categories. Of interest in this manipulation was whether the facilitated relation would be selected more often for the plant category that it was diagnostic of. As the frequency with which the relation appears with the two plant categories is the same, any such effect would suggest that the properties of the concept influence relation selection. Such a demonstration would be important as it would have implications for previous theories (for example the CARIN model has no mechanism by which the diagnosticity of features can influence selection; the frequency of association between relation and concept alone is what is important).

As well as manipulating the diagnosticity of the faciliting feature for plant categories as described above, we also investigated whether diagnostic evidence for conceptual categories would influence interpretation. For example, a particular feature of the beetle concepts was perfectly diagnostic of a relation, whereas a feature of the plant concept was a facilitating feature for that relation. Of interest was whether the beetle or the plant would be more influential in people's selections of relations for beetle-plant compounds.

Half of the participants saw beetle names as the modifier and half of the participants saw plant names as the modifier. The experiment was therefore designed to investigate the interaction between beetle (i.e. beetle feature diagnosticity) and plant (i.e. plant feature diagnosticity and feature facilitation) influence as well as effects of syntax (i.e. whether the beetle or plant is the head or the modifier).

5.1 Method

5.1.1 Participants

Eighteen students of University College Dublin took part in the experiment.

5.1.2 Materials

Training items resembled those of Experiment 1; each depicted a beetle and plant with one of three possible kinds of eating behaviour holding between them, namely "beetle eats from the top of the plant", "beetle eats from the leaf of the plant" and "beetle eats from the trunk of the plant". The "beetle eats from the trunk of the plant" relation required the facilitating property of exposed trunk (i.e. no thorns on the trunk) to be present in the plant. There were also four beetle categories (named BEKEPS, CALARS, DUSUMS and GAMAYS) and two plant categories (named SEEBS and TAUDS) that participants were also required to learn during training. In a first transfer phase, materials were eight noun-noun compounds of these category labels (e.g. SEEB GAMAY; all eight possible pairs of beetle and plant labels were used). This was followed by another transfer phase like the transfer phase used in

Table 5 The 16 items used in the training phase of Experiment 2

	Categories			Beetle features			Plant features		
Item	Rel	Bcat	Pcat	B1	B2	B3	P1	P2	P3
1	1	1	1	3	1	1	1	1	3
2	1	1	1	1	1	1	1	2	1
3	1	1	1	1	1	1	2	1	3
4	3	1	1	1	1	3	1	3	3
5	1	2	1	4	2	1	1	1	2
6	1	2	1	2	2	1	3	3	3
7	1	2	1	2	2	1	1	3	2
8	3	2	1	2	2	3	1	2	3
9	2	3	2	3	2	2	2	1	1
10	2	3	2	3	3	2	2	2	1
11	2	3	2	3	3	2	1	2	2
12	3	3	2	3	3	3	2	3	3
13	2	4	2	4	1	2	2	2	2
14	2	4	2	4	4	2	2	3	1
15	2	4	2	4	4	2	3	2	2
16	3	4	2	4	4	3	2	2	3

Experiment 1: participants were presented with beetle and plant exemplars without any eating behaviour taking place and rated the likelihood of the three relations for the pair. The purpose of this stage was to test how well participants had learned the training items and also to investigate how people responded to the facilitated relation when the facilitating feature was present or absent. There were 32 items presented in this stage in total: the 16 items presented during training plus 16 filler items.

Table 5 gives the abstract category structure of the training items. For each participant, the three abstract beetle dimensions (B1, B2 and B3) were randomly mapped to the concrete physical dimensions of shell colour, shell pattern and facial expression and the two non-facilitating plant dimensions (P1 and P2) were randomly mapped to the concrete physical dimensions of leaf shape and leaf colour. The two independent relations (Relations 1 and 2) were randomly mapped to the concrete relations "beetle eats from the top of the plant" and "beetle eats from the leaf of the plant". The sole abstract facilitating feature (a 3 on P3) and its associated relation (Relation 3) were mapped to "no thorns on trunk" and "beetle eats from the trunk of the plant" respectively. The remaining two possible values on dimension P3 were randomly mapped to two different configurations of thorns on the trunk, which were intended to indicate that eating from the trunk was impossible. The mapping of the four beetle names and the two plant names to the abstract beetle and plant categories was also randomized across participants.

The two plant categories have identical diagnosticity profiles (i.e. the diagnositcity of features on P1, P2 and P3 are the same for the three plant categories) and so are by definition equally easy to learn. The four beetle categories also have identical

diagnosticity profiles. The facilitating feature for Relation 3 (a 3 on dimension P3) occurs seven times in total: four of those are with the four occurrences of Relation 3. A 3 on dimension P3 occurs five times in total for Plant 1 but only two times in total for Plant 2. A question of interest is whether Relation 3 would be preferred for Plant 1 or for Plant 2 (that relation occurs equally often in both plant categories). If Relation 3 is selected equally strongly for both Plant 1 and Plant 2, then this suggests that it is the frequency with which the plant category is associated with the relation that influences the relation selection; however, if Relation 3 is selected more often for Plant 1 than for Plant 2, then this will be evidence that the intrinsic features of the Plant categories are influencing the selected relation.

Features on dimensions B1 and B2 are designed to have high diagnosticity for the four beetle categories. The beetle categories are therefore relatively easy to learn. Dimension B3 is perfectly diagnostic for the relations, making the relations easy to learn also.

5.1.3 Procedure

The training phase consisted of three sub-stages, in which participants learned to distinguish between the plant, beetle and relation categories. During each training sub-phase, the 16 training items were presented to participants sequentially on a web-page in a random order. Underneath each item, participants were presented with a questions of the form "What kind of plant is seen in this picture?", "What type of beetle is seen in this picture?" and "How does this ⟨*Beetle*⟩ eat this ⟨*Plant*⟩?" in the plant learning, beetle learning, and relation learning training sub-stages, respectively. Underneath the question was a series of buttons on which participants could select what they believed to be the correct category. After participants had made their selection, they were given feedback about whether or not their guess on that trial had been correct. Each of the three sub-stages was repeated until participants had correctly classified 75 % or more of the 16 items. Training was followed by the compound transfer phase and then the exemplar transfer phase, in which participants rated the compounds and exemplar items on the scales provided, as in Experiment 1.

5.2 Results

5.2.1 Performance During Training

All but one of the participants successfully completed the training phase. For the remaining 17 participants, successful learning took on average 2.06 iterations of the 16 items for the two plant categories, 4.06 iterations of the items for the four beetle categories, and 3.59 iterations of the items for the three relation categories.

Thus, participants learned to distinguish between the various categories quite quickly, consistent with the fact that the categories were designed to be easy to learn.

5.2.2 Performance During the Compound Transfer Stage

In the compound transfer stage, category order was a between-subject factor: half of the participants saw compounds of the form "⟨*Beetle*⟩ ⟨*Plant*⟩" and half saw compounds of the form "⟨*Plant*⟩ ⟨*Beetle*⟩". Of key interest was whether the training on exemplar items would transfer to relation likelihood ratings for the noun-noun compounds consisting of the learned beetle and plant categories. Previous findings have suggested an asymmetry between the role of the modifier and the head noun in conceptual combination [6] and so we were also interested in whether responses would be affected by the order in which the constituent category names were presented. For example, perhaps it is the concept in the modifier position that is most influential in determining the likelihood of different relations for a compound. Alternatively, perhaps it is the concept in the head position that is most influential.

The four beetle categories used in the experiment can be grouped into two pairs corresponding to how the three relations were associated with the categories during training: Beetle 1 and Beetle 2 were both associated equally strongly with Relation 1 (75 % of the exemplars of these categories were presented with Relation 1) while Beetle 3 and Beetle 4 were both associated equally strongly with Relation 2 (75 % of the exemplars of these categories were presented with Relation 2). In our analysis of the data we were interested in whether the beetle category name used in a compound would influence participants' relation selection. Therefore, in our ANOVA, "beetle category name predicts Relation 1" and "beetle category name predicts Relation 2" were the two levels of a "beetle name influence" factor. The actual beetle category name in the compound was also a (four level) factor in the ANOVA, nested under the beetle name influence factor. The category ordering used in the compounds (i.e. "⟨*Beetle*⟩ ⟨*Plant*⟩" or "⟨*Plant*⟩ ⟨*Beetle*⟩") was a 2-level (between-subject) factor. Response relation (i.e. which of the four relations a rating was for) was also included as a within-subject factor. What is of interest is how participants' ratings of relation likelihood vary depending on the beetle name used in the compounds, the plant name used in the compounds, and the ordering of the beetle and plant names in the compounds: in other words, we are interested in the interaction effects between the response relation factor and the other factors of the ANOVA.

The interaction between the "beetle name influence" factor and response relation was significant ($F(2, 30) = 8.14,\ p < 0.01$); whether the beetle name present in the compound tended to predict Relation 1 or Relation 2 influenced participants' relation selections. However, there was no effect of the beetle category nested within the "beetle name influence" factor. In other words, whether participants saw Beetle 1 or Beetle 2 (associated with Relation 1), or Beetle 3 or Beetle 4 (associated with Relation 2) did not effect their ratings of relation likelihood. The interaction between plant category and response relation was also significant ($F(2, 30) = 4.22,$

$p = 0.02$); the plant category in the compound tended to influence participants' relation selections. The results for the beetle and plant interactions therefore show that participants' learning of the training exemplar items did indeed transfer to their relational responses for the noun-noun compounds.

The interaction between category ordering, beetle name influence and response relation was not significant ($F(2, 30) < 1$); there is no evidence that the influence of beetle category on participants' relation selections differed depending on whether the beetle concept was in the head or modifier position. However, the interaction between category ordering, plant category and response relation was significant ($F(2, 30) = 4.54$, $p = 0.02$); the influence of the plant category name on relation selection differed depending on whether the plant category name was the modifier word or the head word.

To further investigate the response relation's significant interaction with the beetle influence, plant category name, and category ordering factors, the rating data for each of the three relations were analysed separately. For the data for Relation 1, there was a significant main effect of beetle influence ($F(1, 15) = 11.78$, $p < 0.01$); as expected, participants rated Relation 1 more highly when the beetle category was Beetle 1 or Beetle 2 than when the beetle category was Beetle 3 or Beetle 4. There was also a significant main effect of plant category name ($F(1, 15) = 4.84$, $p = 0.04$); Relation 1 was rated more highly for Plant 1 compounds than for Plant 2 compounds. However, the plant category name and category ordering interaction was not significant ($F(1, 15) = 2.50$, $p = 0.14$).

For the Relation 2 ratings, there was a significant main effect of beetle influence ($F(1, 15) = 8.22$, $p = 0.01$), with Relation 2 rated more highly when the beetle category was 3 or 4 than when the beetle category was 1 or 2. The plant category factor was marginally significant ($F(1, 15) = 3.86$, $p = 0.07$). The plant category and category ordering interaction was significant ($F(1, 15) = 7.31$, $p = 0.02$); participants rated Relation 2 higher for Plant 2 than Plant 1 when plant name was in the modifier position but not when plant name was in the head noun position.

For the Relation 3 ratings, there were no significant effects. In particular, there was no main effect of plant category ($F(1, 15) < 1$); ratings for Relation 3 were not affected by the plant category in the compound. This was a comparison of interest in the design of the experiment, as the facilitating feature of Relation 3 occurred more often in Plant 1 exemplars than in Plant 2 exemplars. However, this manipulation appeared to have no affect on participants' ratings of relation likelihood for the compounds. As we hypothesized above, one possible reason for this is that a facilitating feature being true for a concept is a necessary but not sufficient condition for the selection of the corresponding dependent relation, and therefore the dependent relation need not be selected even when the facilitating feature is present.

Conducting separate analyses of variance for the different levels of a factor for which there is a significant interaction as we have done above can be a useful way of investigating the nature of that interaction; however, it does not truly explore the interaction effect, as different levels of the two factors are not considered simultaneously [37, 38]. Since we are interested in whether the relations consistent

with training are in general rated more highly than the alternative relations not consistent with training, an analysis examining more than one level of both factors simultaneously is desirable. For example, we are interested in whether or not the ratings for Relation 1 for Plant 1 and Relation 2 for Plant 2 (i.e. the two relations consistent with those two plant categories during training) are significantly higher than the ratings for Relation 2 for Plant 1 and Relation 1 for Plant 2 (i.e. the two relations that are not consistent with those two plant categories during training). Such an effect involves both levels of both factors.

We therefore conducted a planned contrast for each of the interactions that proved significant in the original repeated measures ANOVA. Each contrast compared the mean of participants' ratings for the relations which were consistent with training against those ratings that were not consistent with training. Each participants' response in each condition of each interaction was averaged. Each contrast was evaluated as a paired t-test, with the consistent-with-training and not-consistent-with-training samples matched by participant. First of all the beetle name influence $\times$ response relation interaction was investigated. The mean ratings of Relation 1 for the "beetle predicts Relation 1" condition and of Relation 2 for the "beetle predicts Relation 2" condition ($M = 2.46$) were significantly higher than the mean ratings of Relation 2 for the "beetle predicts Relation 1" condition and the ratings of Relation 1 for the "beetle predicts Relation 2" ($M = 1.35$; $t(33) = 4.55$, $p < 0.001$); participants clearly learned to identify the relations associated with the beetle categories during the training phase of the experiment and transferred that learning to the compound interpretation task. For the plant category name $\times$ response relation interaction, the mean ratings of Relation 1 for Plant 2 and Relation 2 for Plant 1 ($M = 2.20$) were significantly higher than the mean ratings of Relation 2 for Plant 1 and Relation 1 for Plant 2 ($M = 1.60$; $t(33) = 2.53$, $p = 0.02$). Again, participants clearly learned to identify the relations associated with the plant categories during the training phase of the experiment, and transferred that learning to compound interpretation.

To investigate the category ordering $\times$ plant category name $\times$ response relation interaction, the above planned contrast for plant categories was repeated with the participants in the "plant name is modifier" and the "plant name is head noun" conditions considered separately. For participants that saw the plant in the modifier position, the mean ratings of Relation 1 for Plant 2 and Relation 2 for Plant 1 ($M = 2.56$) were significantly higher than the mean ratings of Relation 2 for Plant 1 and Relation 1 for Plant 2 ($M = 1.27$; $t(15) = 3.35$, $p < 0.01$). Therefore, it seems that participants used the plant category name to guide their ratings of relation likelihood when the plant category name was in the modifier position. However, for participants that saw the plant as the head noun, there was no difference in the mean ratings of Relation 1 for Plant 2 and Relation 2 for Plant 1 ($M = 1.88$) and mean ratings of Relation 2 for Plant 1 and Relation 1 for Plant 2 ($M = 1.90$; $t(17) = 0.07$, $p = 0.95$). Clearly, participants used the plant category name to guide their ratings of relation likelihood only when the plant category name was in the modifier position. This is consistent with the fact that it is more usual in natural language to name animate entities by what they eat than to name inanimate entities by what

they are eaten by (e.g. consider "fruit fly" and "apple maggot") and suggests that participants parsed the two word "⟨*Plant*⟩ ⟨*Beetle*⟩" phrases like natural language compounds.

In summary, the results of the noun-noun compound stage of the experiment show that participants' learning of the relations and their associations with beetle and plant categories during training transferred to a task involving noun-noun compound interpretation. This is important as it demonstrates how the interpretation of compounds can be derived from information about how concept exemplars tend to co-occur together.

5.2.3 Ratings of Exemplar Transfer Items

The second transfer stage required participants to rate relation likelihood for 32 beetle-plant exemplar items (16 test beetle-plant pairs previously seen in training and 16 randomly generated fillers). For the 16 test items, the relation suggested by the beetle category and the relation suggested by the plant category were always the same (for example Beetle 1 and Beetle 2 exemplars always occurred with Plant 1 exemplars, and the relation associated with these categories is Relation 1; see Table 5). The data show that participants successfully learned the relations associated with these 16 items during training: for every one of the 16 items, the highest rated relation was the relation that existed between that pair in training. In particular, the response for the correct relation ($M = 2.94$, $SD = 1.31$) for items was higher than the average response for the other two relations ($M = 1.14$, $SD = 0.88$; collapsing across items, $t(16) = 55.90$, $p < 0.001$, collapsing across subjects, $t(15) = 120.17$, $p < 0.001$).

5.3 Modelling Relation Selection in Compound Interpretation

Our hypothesis about how people decide on likely relations for a compound is that the two lexemes in the compound activate stored memory traces (i.e. exemplars) of the concepts denoted by those lexemes. Exemplars differ in how typical they are for particular conceptual categories and we would expect the likelihood of an exemplar's activation to be proportional to its typicality for the categories named in the compound. As concept instances usually do not happen in isolation but rather in the context of other concepts, this naturally results in extensional relational information about activated exemplars also becoming activated. This activated relational information is then available to form a basis for determining the likely relation for the compound. A strength of this hypothesis is that it incorporates both intensional information about concepts' features (in the form of concept typicality) and also extrinsic, distributional information about how concepts tend to combine (in the form of relational information associated with activated exemplars). In this section, we present a model instantiating this hybrid approach.

In the context of our experiment, the extensional, relational information about beetle and plant exemplars participants held in memory is revealed in how they rated relational likelihood during the exemplar transfer stage of the experiment. For each of the 16 exemplars, we therefore assume that the average ratings for each of the relations describes our participants' knowledge about how exemplars combine with other exemplars. The triad of average ratings for each of the three relations can then be regarded as a vector in a 3-dimensional relation space. We can calculate the relation vector $\mathbf{r}_{B,P}$ for the novel compounds "$B\ P$" or "$P\ B$" as

$$\mathbf{r}_{B,P} = \frac{\sum_{e \in U} (typ(e_b, B) + typ(e_p, P))^{\alpha} \cdot \mathbf{r}_e}{\sum_{e \in U} (typ(e_b, B) + typ(e_p, P))^{\alpha}} \tag{6}$$

where e denotes one of the 16 beetle-plant exemplar items rated in the exemplar transfer stage, $typ(e_b, B)$ denotes the typicality of the beetle exemplar present in item e in beetle category B and $typ(e_p, P)$ denotes the typicality of the plant exemplar present in item e in plant category P. U is the set of 16 beetle-plant exemplar pairs and α is a magnification parameter to be estimated empirically which describes the relative importance of exemplar typicality. This model is a specific implementation of a broader, exemplar-based model of conceptual combination that we have proposed elsewhere and have successfully used to model relation selection for natural language compounds [3, 39, 40].

In this model, we require a measure of how typical of a conceptual category particular exemplars are.We use the probability scores produced by the GCM as a means for computing concept typicality (although other methods for measuring typicality could have been used). In computing the GCM typicality ratings, we set c, the GCM magnification parameter, to 1. The attentional weights for the three plant dimensions and the three beetle dimensions were not treated as free parameters but were estimated using an information-theoretic method [40]. Therefore, we utilize only one free parameter, α, in fitting our model to the compound rating data.

We compared the relation vector outputted by the model for the eight possible compounds to the relation vectors derived from participants' ratings in the compound transfer phase of the experiment. Across all eight compounds and three relations, the agreement between the model and the data was high ($r = 0.85$, $p < 0.001$ with optimal $\alpha = 12$). Looking at the three relations separately revealed high correlations for Relation 1 ($r = 0.95$, $p < 0.001$) and Relation 2 ($r = 0.85$, $p < 0.001$); however, the correlation for Relation 3 was not significant ($p = 0.27$). The model does not match the data for Relation 3 because participants' responses for Relation 3 across the eight items are essentially the same (with some small random variability), consistent with the fact that the four beetle categories and two plant categories occur equally often with that relation. Similarly, the model gives values for Relation 3 which are almost identical across the eight items.

The evaluation of the model above uses relation vectors which are based on the actual association of relations with exemplars that participants learned, as described by participants' responses to the three relations for the 16 items presented during the exemplar transfer stage. We also evaluated the model using the association between the exemplar and the relation that was specified in the abstract category structure.[4] In this version of the model, the relation exemplars associated with conceptual exemplars are not based on data obtained from participants in the exemplar transfer stage. However, there was still high agreement between the model's predications and the data ($r = 0.77$, $p < 0.001$, with $\alpha = 11$).

The modelling results are important as they demonstrate that a model which is based solely on information about exemplars, and about the relational links between those exemplars and other exemplars they co-occurred with, can explain how people determine the correct or most likely relations for noun-noun compounds.

6 Experiment 3

Experiment 2 established that participants can use information about learned exemplars to determine the correct interpretations of compounds. In Experiment 3, we aim to investigate three issues that may be important in determining the most appropriate interpretation for a compound. Firstly, the experiment aims to investigate the influence of properties of concepts not captured purely by the abstract category structure or by the frequency with which the concepts appear with different relations. For example, if the two concepts referenced in a compound are identical with respect to the complexity of their representation, how well they are associated with various alternative relations (and so on), but are of differing levels of animacy, we might expect the relation associated with the more animate concept to be selected by participants more often than a different relation associated equally strongly with the less animate concept. In the experiment, all three relations again involve a beetle eating a plant. Since in each case the beetle is the agent of the scenario, it is possible that the semantics of the beetle concepts might be more relevant to relation selection than the semantics of the plant concepts.

Secondly, the experiment again manipulates the ordering of the two nouns within the compound: given two categories named A and B, our experiment investigates whether the compound "$A\ B$" is interpreted in the same way as the compound "$B\ A$". Furthermore, of interest was whether the location of the more animate concept in the compound would have an effect on interpretation. For example, since the combined concept is an instance of the head concept, we might hypothesize that compounds for which the head concept is more animate than the modifier concept may be easier to interpret correctly.

[4]For example, the relation vector for exemplars occurring with Relation 1 (i.e. the beetle and plant exemplars in items 1, 2, 3, 5, 6 and 7) would be [1, 0, 0].

Table 6 The abstract category structure of Experiment 3

Learn	*Trans.*	Nr	Rel	Bcat	Pcat	B1	B2	B3	P1	P2	P3
l		1	1	1	3	4	1	1	3	2	3
l		2	1	1	3	4	4	1	2	3	3
l	*t*	3	1	1	3	1	1	1	3	3	2
l	*t*	4	1	1	3	4	1	2	3	3	3
l	*t*	5	2	2	2	2	2	2	2	2	3
l		6	2	2	2	2	2	1	2	3	2
l		7	2	2	2	2	3	2	2	2	1
l	*t*	8	2	2	2	2	2	3	2	2	2
l	*t*	9	3	3	1	3	3	3	4	1	2
l	*t*	10	3	3	1	3	3	2	1	1	1
l		11	3	3	1	2	3	3	4	4	1
l		12	3	3	1	3	2	3	4	1	1
l	*t*	13	1	4	4	1	1	4	4	4	4
l	*t*	14	2	4	4	4	1	4	4	1	4
l	*t*	15	3	4	4	4	4	4	1	1	4
	t	16	–	1	1	4	1	1	4	1	1
	t	17	–	3	3	3	3	3	3	3	3
	t	18	–	2	4	2	2	2	4	1	4
	t	19	–	4	2	4	1	4	2	2	2

Finally, were interested in the effect of concept similarity: would compounds consisting of similar constituent categories tend to be interpreted in similar ways. In our previous work on the influence of conceptual similarity on the interpretation process [3] we utilised the IIC metric [41], which uses WordNet as a measure of the similarity between various concepts. However, with laboratory-generated categories we can precisely define and control the similarity between the concepts in terms exemplar distance.

6.1 Method

6.1.1 Participants

The participants were 42 university students.

6.1.2 Materials

Table 6 presents the abstract category structure for Experiment 3. There are 19 items in total; the first and second columns in the table indicate if the item was one of the

15 items used in the learning phase of the experiment (*l*) or as one of the 13 items used in the transfer stage of the experiment (*t*). There were four beetle categories (Bcat), four plant categories (Pcat) and three relation categories (defined by features instantiated on dimensions (B1, B2, B3, P1, P2 and P3). Unlike Experiment 2, the beetle and plant categories had identical structure (for example, the four exemplars of Pcat1 have the same structure as the four exemplars of Bcat1).

Beetles and plants were associated with particular relations; Bcat1, Bcat2 and Bcat3 were associated with Relations 1, 2 and 3, respectively, whereas Pcat1, Pcat2 and Pcat3 were associated with Relations 3, 2 and 1, respectively. Bcat4 and Pcat4 were not associated with any relations; the three exemplar instances of these categories in the learning phase appeared once with each of the three relations. The features of beetles and plants were sometimes diagnostic of a concept category (e.g. a feature associated with Bcat1 is a 1 on dimension B3: 3 of the 4 Bcat1 training exemplars have a 1 on dimension B3 while only one of the remaining 11 training exemplars do). Also, the intrinsic features of beetles and plants are sometimes diagnostic of a relation category (values on dimensions B1, P1, B2 and P2 are quite diagnostic of relations).

By holding beetle and plant category structure identical, it was hoped that aspects of conceptual combination that were independent of the feature and relational distributional structure learned for the constituent concepts could be investigated. So, for example, for the compound "Bcat1 Pcat1", there is equal evidence for Relation 1 (given by the beetle category) and Relation 3 (given by the plant category). We were interested in whether Relation 1 would be rated equally likely as Relation 2. If not, we were interested in whether this inequality was due to the ordering of the items (for example, the relation associated with the modifier concept might be preferred) or the saliency of the constituent concepts (for example, the relation associated with the beetle category might be preferred).

The beetle and plant categories were also designed to differ in terms of their similarity. For example, categories Bcat1 and Bcat4 are more similar to each other than Bcat3 and Bcat4 are: the features for Bcat1 and Bcat4 overlap to a greater extent than the features for Bcat3 and Bcat4 do. The aim of varying categories with respect to their similarity was to investigate whether similar categories would yield similar patterns of relation likelihood ratings. In particular, Bcat4 (and Pcat4) occurs equally often with the three relations; therefore if category similarity has no effect we would expect people to select each of the relations equally often for this category. However, if similarity influences participants' relation selection, then we would expect that Relation 1 would be selected more often than Relations 2 or 3.

As in Experiment 2, the abstract category structure was randomly mapped to concrete features in a way that was unique for each participant. The three concrete beetle features, three concrete plant features, and three concrete relations were the same as in Experiment 2.

6.1.3 Procedure

The training phase, compound transfer stage and exemplar transfer stage procedures were identical to Experiment 2. The compound transfer stage followed the exemplar transfer stage. In the exemplar transfer stage, participants were presented with 13 beetle-plant items, some of which had appeared in training and some of which were new items (Table 6). The materials used in the compound transfer stage were the 16 possible noun-noun compounds consisting of beetle and plant names.

6.2 Results

6.2.1 Performance During Training

Two of the participants failed to complete the training phase. For the remaining 40 participants, successful learning took on average 5.8 iterations of the training items for the plant categories, 3.9 iterations for the beetle categories, and 2.1 iterations for the relation categories. The participants therefore learned to distinguish between the categories quite quickly, consistent with the fact that the categories were designed to be easy to learn.

6.2.2 Performance During the Exemplar Transfer Stage

Participants' mean ratings of relation likelihood for the nine previously seen exemplar items are presented in Fig. 3 (items 3–15). For each of these items there was a correct relation, namely the one that the item was associated with during training. The difference between the mean response for the correct relation ($M = 2.76$) and the mean response for the two incorrect relations ($M = 1.42$) was significant ($t_s(39) = 7.50$, $p < 0.01$; $t_i(8) = 4.07$, $p < 0.01$), indicating that participants learned which relations tended to co-occur with the items in the training phase.

Participants' mean ratings of relation likelihood for the four exemplar items not previously seen in training are also presented in Fig. 3 (items 16–19). Each of these four items consisted of a prototypical example of each of the four beetle categories and each of the four plant categories (with each beetle and plant category appearing once; see Table 6). For these four items the relation consistent with the beetle exemplar was always different to the relation suggested by the plant exemplar. For each trial, one relation is consistent with the beetle exemplar (r_b), one is consistent with the plant exemplar (r_p) and one is neutral (r_n). One-way repeated measures ANOVAs with response type (r_b, r_p or r_n) as a fixed factor revealed a significant effect of response type ($F_s(2, 39) = 19.10$, $p < 0.01$; $F_i(2, 3) = 24.14$, $p < 0.01$). Pairwise differences between the three response types were investigated using

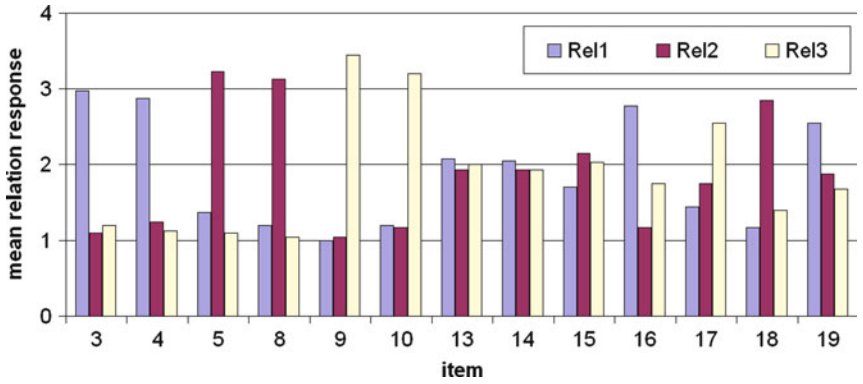

Fig. 3 Participants' mean responses for the exemplar transfer items

planned comparisons. The difference between participants' mean response for the relation associated with the beetle exemplar, r_b ($M = 2.68$), and their mean response for the neutral relation, r_n ($M = 1.44$) was significant ($t_s(39) = 5.63$, $p < 0.001$; $t_i(3) = 5.34$, $p = 0.01$). These results suggest that participants were strongly influenced by the beetle exemplar when making their category judgments. However, the difference between participants' mean response for the relation associated with the plant exemplar, r_p ($M = 1.62$), and their mean response for the neutral relation was not significant ($t_s(39) = 1.11$, $p = 0.27$; $t_i(3) = 0.97$, $p = 0.40$). These results suggest that participants were not influenced by the plant exemplar when judging relation likelihood. Since the beetle and plant categories have identical abstract structure, these results suggest that other factors (such as the animacy of a concept or the role it plays in the relation) are important to interpretation.

To investigate possible effects of category similarity, the data from all 13 items were analysed with a repeated measures ANOVA with beetle category and response relation as within-subject factors and subject as a random factor. There was a significant effect of the category that the beetle exemplar belonged to on participants' responses for the three relations (the interaction between beetle category and response relation was significant; $F(6, 39) = 26.83$, $p < 0.01$). Planned pairwise comparisons (paired t-tests) were conducted to investigate how ratings for the correct relation (i.e. the relation consistent with training) differed for the ratings for the other two relations. For Bcat1, Bcat2 and Bcat3, the ratings for the relation consistent with learning was higher than the two alternative relations ($p < 0.01$ in all cases). However, for the Bcat4 items, there was no evidence that participants we more likely to rate Relation 1 ($M = 2.09$) higher than either Relation 2 ($M = 1.97$; $t(39) = 0.54$, $p = 0.59$) or Relation 3 ($M = 1.91$; $t(39) = 0.69$, $p > 0.50$). Though the difference is in the direction predicted by Bcat4's similarity to Bcat1, there is no evidence that participants made use of Bcat4's similarity to Bcat1 when rating relation likelihood for Bcat4.

In summary, the results suggest that participants were capable of learning the training items, and participants appeared to be influenced by the beetle exemplar but not the plant exemplar.

6.2.3 Performance on the Noun-Noun Compound Transfer Stage

In the noun-noun compound transfer stage, participants rated relation likelihood for each of the 16 compounds that could be formed from combinations of the beetle and plant category names. Half of the participants saw the compounds with beetle in the modifier position and plant in the head position whilst the other half saw the reverse. Again, we were interested in whether or not the training on exemplar items would transfer to the compounds and whether or not participants' responses would be affected by the order of the category labels in the compound.

A 4×4×3×2 repeated measures ANOVA with beetle category, plant category and response relation as within subject factors and category label ordering as a between subject factor was used to analyze the data. The interaction between beetle category and response relation was significant ($F(6, 38) = 59.79,\ p < 0.001$); the beetle category present in the compound influenced relation selections. The interaction between plant category and response relation was weaker, but still significant ($F(6, 38) = 5.35,\ p < 0.01$).Training on exemplar items therefore transferred to the noun-noun compounds. However, there were no other significant interactions found. In particular, the interaction between category ordering, beetle category and response relation was not significant ($F(6, 38) = 1.82,\ p = 0.09$); there is no evidence that the influence of beetle category on relation selections when the beetle was in the modifier position differed from the influence of beetle category on relation selections when the beetle was in the head-noun position. Similarly, the interaction between category ordering, plant category and response relation was not significant ($F(6, 38) < 1$); the influence of the plant category on relation selection did not differ depending on the location of the plant category in the compound.

Planned pairwise comparisons (paired t-tests) were used to investigate the significant interactions further: for Bcat1, Bcat2 and Bcat3, the ratings for the relation consistent with learning was significantly higher than the two alternative relations ($p < 0.001$ in all cases). However, for Bcat4, there were no significant differences between the ratings for the three relations ($p > 0.31$ for each of the three comparisons). For the plants, however, the only significant differences were between the response for Relation 1 and Relation 2 for Pcat2 ($t(39) = 2.12,\ p = 0.04$) and between Relation 2 and Relation 3 for Pcat2 ($t(39) = 3.08,\ p = 0.004$), although the differences for Pcat1 and Pcat3 are also in the expected direction.

In summary, the results of the noun-noun compound stage of the experiment show that participants' learning of the relations and their associations with beetle and plant categories during training transferred to a task involving noun-noun compound interpretation. This is important as it demonstrates how the interpretation of compounds can be derived from information about how concept exemplars tend to co-occur together.

6.3 Modelling Relation Selection

People's judgment of relation likelihood in the compound rating stage of Experiment 3 was modelled in the same manner as for Experiment 2. Again, we evaluate the model in two ways; the first method uses the data from the nine previously seen items in the exemplar rating stage of the experiment as the representation of people's knowledge about the concept instances and how those instances tend to co-occur with other concept instances, while in the second method we use the actual category structure people were exposed to as the representation of exemplar knowledge. The agreement between the model and the data was high across the three relations ($r = 0.87$, $p < 0.001$, with optimal $\alpha = 5$). Looking at the three relations separately revealed a high correlation for Relations 1 ($r = 0.84$, $p < 0.001$), 2 ($r = 0.90$, $p < 0.001$) and 3 ($r = 0.88$, $p < 0.001$). In the version of the model using the actual association between the exemplar and the relation that was specified in the abstract category structure there is still a high degree of agreement between the model's predications and the data ($r = 0.80$, $p < 0.001$, with $\alpha = 3$). Again, the success of the model demonstrates quite convincingly that a model which is based solely on information about exemplars can explain how people determine the correct or most likely relations for noun-noun compounds.

7 Conclusions

The empirical findings we have described in this chapter have several important implications. Firstly, the findings have implications for relation-based theories. In particular, the finding that only beetle exemplars tended to influence relation selection in Experiment 3 (Fig. 3) suggests that factors other than relation frequency are relevant to the interpretation process (since the beetle and plants in our experiment were identical in their degree of association with relations). Complex interactions between concepts and relations (e.g. agency in the EATS(AGENT,OBJECT) relation) is information that is not possible to capture using a taxonomic approach to relation meaning.

Secondly, the fact that participants could learn to identify the relations between exemplars and also transfer that knowledge to a task involving compounds has implications for concept-based theories of compound comprehension. No concept-based theory of conceptual combination has ever adopted an exemplar approach to concept meaning; models based on concept-focused theories tend to represent concepts as frames or lists of predicates. Our approach suggests an exemplar representation is a viable alternative. Also, distributional knowledge about relations forms a natural component of an exemplar representation of concepts, as different concept instances will occur with instances of other concepts with varying degrees of frequency. Given the success of our model, assuming an exemplar representation of concept semantics would seen to offer a natural way of incorporating both

information about concept features and information about relation distribution into a single theory.

Our exemplar-based approach to representing the relations used in conceptual combination is also consistent with recent evidence suggesting that relations are independent semantic representations that become active during the interpretation process. For example, the semantic relations use in compound interpretation can be primed, indicating that relations are semantic structures that exist independently of the representations of the constituent concepts [42, 43]. These findings suggests that relations are themselves complex semantic representations, like the concepts that they link. By modelling both concepts and relations using exemplar-based category structures, our approach provides a framework for modelling the semantic complexity of relations, and provides an account for how intrinsic information (i.e. feature diagnosticity and facilitating features) and extensional, distributional information (i.e. the co-occurrence of concepts with relations) about relations influence which relation is selected for a compound during the interpretation process.

Acknowledgements This research was funded by Irish Research Council for Science, Engineering and Technology Grant RS/2002/758-2 to BD.

References

1. Clark, E. V., & Barron, B. J. (1988). A thrower-button or a button-thrower? Children's judgments of grammatical and ungrammatical compound nouns. *Linguistics, 26*, 3–19.
2. Cannon, G. H. (1987). *Historical change and English word formation*. New York: Lang.
3. Devereux, B., & Costello, F. J. (2006). Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. In R. Sun & N. Miyake (Eds.), *Proceedings of the Twenty-Eigth Annual Conference of the Cognitive Science Society* (pp. 184–189). Mahwah: Cognitive Science Society/Lawrence Erlbaum. ISBN 0-9768318-2-1.
4. Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic.
5. Downing, P. (1977). On the creation and use of English compound nouns. *Language, 53*(4), 810–842.
6. Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 71–78.
7. Kim, S. N., & Baldwin, T. (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*. Sydney: Association for Computational Linguistics.
8. Barr, R. A., & Caplan, L. J. (1987). Category representations and their implications for category structure. *Memory and Cognition, 15*, 397–418.
9. Anggoro, F., Gentner, D., & Klibanoff, R. (2005). How to go from nest to home: Children's learning of relational categories. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (pp. 133–138). Mahwah: Cognitive Science Society/Lawrence Erlbaum. ISBN 0-9768318-1-3.
10. Wisniewski, E. J. (1997). When concepts combine. *Psychonomic Bulletin & Review, 4*, 167–183.

11. Costello, F. J. & Keane, M. T. (2000). Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science, 24*(2), 299–349.
12. Wisniewski, E. J., & Murphy, G. L. (2005). Frequency of relation type as a determinant of conceptual combination: A reanalysis. *Journal of Experimental Psychology: Learning Memory and Cognition, 31*, 169–174.
13. Gagné, C. L., & Spalding, T. L. (2006). Relation availability was not confounded with familiarity or plausibility in gagné and shoben (1997): Comment on wisniewski and murphy (2005). *Journal of experimental psychology. Learning, memory, and cognition, 32*(6), 1431–1437; discussion 1438–1442. ISSN 02787393.
14. Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75*(2), 195–231.
15. Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences, 5*(6), 244–252.
16. Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review, 9*, 542–549.
17. Pexman, P. M., Holyk, G. G., & Monfils, M. -H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition, 31*, 842–855.
18. Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009). Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research, 1282*, 95–102. ISSN 1872-6240.
19. Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238.
20. Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104–114.
21. Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.
22. Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 87–108.
23. Gentner, D. (1981). Verb semantic structures in memory for sentences: Evidence for componential representation. *Cognitive Psychology, 13*, 56–83.
24. Duncker, K. (1945). On problem solving (L. S. Lees, Trans.). *Psychological Monographs, 58*(5, Whole No. 270), 1–113.
25. Keane, M. T. (1985). Ton drawing analogies when solving problems. *British Journal of Psychology, 76*, 449–458.
26. Costello, F. J. (2000). An exemplar model of classification in simple and combined categories. In L. R. Gleitman & J. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Mahwah: Lawrence Erlbaum Associates.
27. Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal Of Experimental Psychology: Learning Memory And Cognition, 26*, 3–27.
28. Costello, F. J. (2001). A computational model of categorisation and category combination: Identifying diseases and new disease combinations. In J. D. Moore & K. Stenning, (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society, University of Edinburgh* (pp. 238–243). Mahwah: Lawrence Erlbaum Associates.
29. Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 924–940.
30. Zaki, S. R., Nosofsky, R. M., Stanton, R. D., & Cohen, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Learning and Memory, 29*, 1160–1173.

31. Yamauchi, T., & Markman, A. B. (2000). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. *Memory & Cognition, 28* 64–78.
32. Yamauchi, T., & Yu, N.-Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition, 36*, 544–553.
33. Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2004). *TiMBL: Tilburg memory based learner, Version 5.1, Reference guide*, Tilburg University (ILK Techn. Rep. 04–02).
34. Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science, 237*, 1317–1323.
35. Nosofsky, R. M. (1985). Overall similarity and the identifcation of separable-dimension stimuli: A choice model analysis. *Perception and Psychophysics, 38*, 415–432.
36. Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classiffication, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700–708.
37. Keselman, H. J. (1998). Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. *Psychophysiology, 35*, 470–478.
38. Boik, R. J. (1987). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics, 18*, 1–40.
39. Devereux, B., & Costello, F. J. (2005). Representing and modelling the meaning of noun-noun compounds. In K. Opwis & I. Penner (Eds.), *Proceedings of KogWis05: the Seventh Biannual Meeting of the German Cognitive Science Society* (pp. 33–38). Basel: German Cognitive Science Society/Schwabe. ISBN ISBN 0-9768318-1-3.
40. Devereux, B. (2007). *The Role of Relational and Conceptual Knowledge in the Interpretation of Noun-Noun Compounds*. Ph.D. thesis, University College Dublin.
41. Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in wordnet. In *Proceedings of ECAI 2004, the 16th European Conference on Artificial Intelligence*, (pp. 1089–1090). Valencia: IOS Press.
42. Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language, 60*, 20–35.
43. Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language, 55*, 89–101.

Child Acquisition of Multiword Verbs: A Computational Investigation

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson

Abstract Traditional theories of grammar, as well as computational modelling of language acquisition, have focused either on aspects of word learning, or grammar learning. Work on intermediate linguistic constructions (the area between words and combinatory grammar rules) has been very limited. Although recent usage-based theories of language learning emphasize the role of multiword constructions, much remains to be explored concerning the precise computational mechanisms that underlie how children learn to identify and interpret different types of multiword lexemes. The goal of the current study is to bring in ideas from computational linguistics on the topic of identifying multiword lexemes, and to explore whether these ideas can be extended in a natural way to the domain of child language acquisition. We take a first step toward computational modelling of the acquisition of a widely-documented class of multiword verbs, such as take the train and give a kiss, that children must master early in language learning. Specifically, we show that simple statistics based on the linguistic properties of these multiword verbs are informative for identifying them in a corpus of child-directed utterances. We present preliminary experiments demonstrating that such statistics can be used within a word learning model to learn associations between meanings and sequences of words.

A. Nematzadeh (✉)
Department of Computer Science, University of Toronto, Toronto, Canada
e-mail: aida@cs.toronto.edu

A. Fazly
School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O.Box 19395-5746, Tehran, Iran
e-mail: afsaneh.fazly@gmail.com

S. Stevenson
Department of Computer Science, University of Toronto, Toronto, Canada
e-mail: suzanne@cs.toronto.edu

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_9,

1 Introduction

Traditional theories of grammar distinguish between lexical knowledge (the individual words that a speaker knows) and grammatical knowledge (the rules for combining words into meaningful utterances). However, there is a rich range of linguistic phenomena in the less explored area between words and combinatory rules/constraints. For example, a multiword lexeme such as *take the train* has an idiosyncratic semantics ("use a train as mode of transport") that suggests its treatment as a lexical unit, since the meaning cannot be compositionally derived in a general manner.[1] But *take the train* also behaves as a syntactic phrase, undergoing various alternative means of expression (e.g., *took a train*, *take the fast train*, *take trains all over Europe*). Much research on language has thus focused on a range of multiword lexemes such as idioms, light verb constructions, noun compounds, and collocations (e.g., [15, 20, 22, 46, 48, 55, 76]). Psycholinguists have also shown the importance of co-occurrence and contingent frequency effects between words, and between words and syntactic patterns in the learning and processing of language (e.g., [5, 57, 70, 72]).

In theories of language acquisition in particular, especially usage-based accounts of language learning (which eschew complex innate linguistic knowledge), the role of multiword constructions has been emphasized (e.g., [40, 41, 74]). However, *computational modelling* of language acquisition has continued to focus on various aspects of word learning (e.g., [33, 37, 49, 65, 77]), or grammar learning (e.g., [17, 69]). Work on intermediate constructions has mostly been limited to identifying general properties of verb argument usages (e.g., [3, 4, 13, 18, 23, 61, 63]), rather than on multiword lexemes. Recent work by Borensztajn et al. [9] uses a probabilistic model (in the DOP framework) to show that a grammar learner can progress from highly lexicalized to multiword tree fragments, on the basis of statistical patterns in the kind of input children receive. Bannard and Matthews [7] further give evidence from human subjects that children are sensitive to the frequencies of multiword sequences. These studies provide evidence that children recognize and produce certain (e.g., high-frequency) multiword sequences in their input, but do not address what sort of cues (other than, e.g., frequency) a child might use to identify, and treat differentially, the various distinguished types of multiword lexemes suggested by linguistic analyses.

Thus in the study of child language acquisition, much remains to be explored concerning the precise computational mechanisms that underlie how children learn to identify different types of multiword lexemes—that is, how they recognize that an idiosyncratic semantics is associated with a sequence of words (rather than single words plus combinatory rules), and how the idiosyncratic meaning relates to the surface (lexical and syntactic) form of a particular combination. In contrast, there

[1] A compositional approach to *take the train* would depend on knowledge of a very specialized meaning of *take* restricted to occur with a narrow range of objects, which is essentially an alternative lexicalization of the necessary knowledge. See Fazly et al. [31] for a computational approach to the restricted productivity of such expressions.

has been significant work in computational linguistics on this very topic, with the development of statistical measures, both for identifying multiword lexemes in a corpus, and for determining the syntactic and semantic behaviour of the particular type of multiword lexeme in question (e.g., [8,19,21,25,28,30,43,50,53,67,71,75]). The goal of our research here is to explore whether this computational work on multiword lexemes can be extended in a natural way to the domain of child language acquisition, where an informative cognitive model must take into account the two issues of what kind of data the child is exposed to, and what kinds of processing of that data is cognitively plausible for a child.

In pursuing these questions, we focus in particular on the acquisition of multiword verbs, such as *take the train* and *give a kiss*. These constructions are a rich and productive source of predication which children must master in most languages, doing so at very young ages [41]. For example, consider the following conversation from the CHILDES database ([11], sarah130a.cha):

*MOT: you're not gonna *take any toys* down to the beach today you know.
*CHI: why?
...
*MOT: we have to *take the train*.

Here, the mother uses the verb *take* first in its core literal meaning (in *take any toys*), and then within a multiword lexeme in which *take* has a non-literal meaning and combines with the particular argument to express the use of a mode of transportation (in *take the train*). The child's further responses within this conversation give no indication that she is puzzled by these very different usages of *take*. Yet they do pose a very significant puzzle for researchers: It has been noted that children learn highly frequent verbs (such as *take*) first (e.g., [41]), and yet it is precisely these verbs that are also the most polysemous, showing a wide range of metaphorical sense extensions in multiword lexemes, which children recognize and deal with effectively [16,44,73].

Research over the last few years has shown that the distinctions among literal and non-literal verb–argument combinations (such as *take the toys* versus *take the train* or *take a nap*) are in principle learnable based on statistics over usages of such expressions (e.g., [30,75]). However, such work depends on very large amounts of data (from corpora on the order of 100 M words) and on sophisticated statistical and grammatical calculations over such data. The goal here is to determine what is learnable through the means available to a child—that is, on the basis of data in child-directed speech and using simpler, cognitively plausible calculations.

We begin by summarizing the motivation and approach to deriving simple statistics based on the linguistic properties of the multiword lexemes under study (first presented in [32]). We then present new experiments that show that such statistics can be informative in identifying such multiword lexemes in child-directed speech. Then we turn to a novel approach for incorporating these statistical measures into an existing model of word learning, to show further that such statistics can be used within a natural process of word learning to associate a single meaning with a sequence of words. In this way, we take a first step toward computational

modelling of acquisition of the kinds of multiword verbs that children must master early in language learning, shedding light on the mechanisms that could underlie a usage-based model of this process.

2 Multiword Lexemes with Basic Verbs

The highly frequent and highly polysemous verbs referred to above include what are called "basic" verbs—those that express physical actions or states central to human experience, such as *give*, *get*, *take*, *put*, *see*, and *stand*, among others. These verbs undergo metaphorical sense extensions of their core physical meanings that enable them to combine with various arguments to form multiword lexemes [15,58,59,62]. We focus here on expressions in which a basic verb is combined with a noun in its direct object position to form either a literal combination (as in *take the toys*) or a multiword lexeme (such as *take the train*, *take a nap*). We refer to all such expressions (both literal and non-literal) as verb–noun combinations or verb–noun pairs, with the understanding that the verb is a basic verb.

Verb–noun combinations that form multiword lexemes are very frequent in many languages (e.g., [1, 20, 45, 46, 51, 54]). Such expressions show a range of semantic idiosyncrasy, where the semantics of the multiword lexeme is more or less related to the semantics of the verb and the noun separately [38, 66]. Thus, verb–noun combinations can be viewed as lying on a continuum (without completely clear boundaries) from entirely literal and compositional, to highly idiomatic. However, for convenience we can think of classes of constructions on this continuum, each identified by a particular way in which the verb and the noun component contribute to the meaning of the construction. Following [30], we consider four possible classes; these are listed below with an example from the child-directed speech used in our experiments along with some information about the semantic contribution of the components of expressions in that class:

1. **Literal combination or LIT**
   - *Give (me) the lion*
     - *Give*: physical transfer of possession
     - *Lion*: a physical entity
2. **Abstract combination or ABS**
   - *Give (her) time*
     - *Give*: abstract transfer or allocation
     - *Time*: an abstract meaning
3. **Light verb construction or LVC**
   - *Give (the doll) a bath*
     - *Give*: convey/conduct an action
     - *Bath*: a predicative meaning

4. **Idiomatic combination or IDM**
   - *Give (me) the slip*
     - *Give*, *slip*: no/highly abstract contribution

These classes are important in the context of child language acquisition because there is a clear connection between the linguistic properties of each class and the meaning of the expressions in the class. Such a relation can enable language learners to generalize their item-specific knowledge, for example by making predictions about the meanings of new expressions based on their likely class. For example, when a child hears a new expression such as *give a shout*, if she recognizes that this is likely an LVC, then she can infer that it roughly means the same thing as the noun—i.e., *shout*—which contributes the predicative meaning, and also infer any other properties holding of LVCs more generally.[2]

The four classes of expressions above have differing linguistic behaviours that can be cues to the underlying distinctions among the classes [30]. Specifically, expressions from each class exhibit particular lexical and syntactic behaviour that closely relate to the semantic properties of the class. We next elaborate on these properties and behaviours, and describe how they can form the basis for statistical measures for distinguishing the classes.

3 Linguistic Properties and the Usage-Based Measures

It has been shown that children are sensitive to the frequency of occurrence of multiword sequences (e.g., [7]). However, simple co-occurrence frequency of a verb and a noun (or measures of association between the two) do not suffice for accurate identification of multiword verb–noun lexemes [29]. We thus further hypothesize that children are also sensitive to the syntactic and semantic properties of each class of verb–noun combination. As a first step to examining this hypothesis, we need to verify whether information about such properties is available in the input children receive, and whether the available information is useful for determining the semantic class of a given combination. We note that there is some overlap in the properties exhibited by the various non-literal classes. We thus further simplify our task here by aiming to distinguish the non-literal expressions (those from ABS, LVC, IDM) from literal ones (LIT). There is only one instance of an IDM in our data, hence in our presentation of the measures here, we discuss the properties with respect to the ABS+LVC classes.

As noted earlier, computational linguistic studies have developed sophisticated statistical measures based on such properties, which have achieved success in identifying non-literal combinations when evaluated on large amounts of text corpus

[2]For example, adult competence with the language includes the knowledge that this refers to a single occurrence of a bounded ‘shouting’ action [12, 76].

data (e.g., [28, 30]). Given the hypothesized importance of simplicity in language learning (c.f. [60]), our goal here is to use simpler measures (tapping into similar properties) that are more cognitively plausible, and that are robust when used with smaller amounts of child-directed speech (CDS). We note that some of the measures explained in this section are taken and adapted for this purpose from Fazly [29]. The resulting measures fit into three groups based on the linguistic properties of the verb and the noun in a verb–noun combination: the degree of association of the verb and noun, the semantic properties of the noun, and the degree of syntactic fixedness of the expression.

3.1 Association of a Verb–Noun Pair

In a literal verb–noun combination, where the verb contributes its core physical semantics, a wide variety of nouns can occur as the noun component (e.g., one can *give an apple*, *a book*, *a car*, *a dog*, etc.). In contrast, in a non-literal combination, the verb has an abstract and/or metaphorical meaning and hence can combine with a set of nouns that is semantically, and somewhat idiosyncratically, restricted (e.g., *give a groan/cry/yell*, but not *give a gripe*, [31]). Moreover, the latter group of nouns often contribute a specific abstract meaning to the combinations they appear in, and hence may not occur as the direct object of other verbs as frequently as do concrete nouns. As a result, we expect the verb and the noun component in non-literal expressions to co-occur more often compared to the components of literal combinations [14, 27]. Below we explain two different measures capturing the marked frequency of a verb-noun pair.

Frequency. The simplest way to measure the association of a verb and a noun is by the frequency of co-occurrence of the verb–noun pair $\langle v, n \rangle$, as in:

$$\text{Cooc}(v, n) \doteq \text{freq}(v, n, \text{gr} = \text{dobj}) \tag{1}$$

where gr = dobj indicates that the noun is the direct object of the verb. We assume that children are able to keep track of simple counts of such verb–noun pairs.

Conditional Probability. Although non-literal expressions are expected to co-occur more often compared to literal expressions, the co-occurrence of some literal expressions is also significant (e.g., *take the toy* in child-directed speech). However, the noun in a non-literal expression generally does not occur with as diverse a set of verbs as a noun in a literal expression. For example, *apple* can be used in many literal expressions with different verbs: *give the apple*, *take the apple*, *eat the apple*, and *wash the apple*, whereas *decision* only occurs in one non-literal verb–noun

combination: *make a decision*.[3] In other words, while the verb in a LIT expression is typically thought of as selecting for a noun in direct object position, in a non-literal expression the noun can be viewed as selecting for a verb (e.g., [24,43]). We measure this property by computing the conditional probability of a verb–noun pair given the noun (CProb).

$$\begin{aligned} \text{CProb}(v,n) &\doteq P(v|n, \text{gr} = \text{dobj}) \\ &= \frac{\text{freq}(v,n,\text{gr} = \text{dobj})}{\sum_{v'} \text{freq}(v',n,\text{gr} = \text{dobj})} \end{aligned} \tag{2}$$

This measure is still a very simple one for children, since it is composed of two frequency counts, although we should note that it does assume that children are able to keep track of the count of a noun as the direct object of any verb.[4]

3.2 Semantic Properties of the Noun

There is evidence that children are sensitive to the semantic differences between the nouns in a literal versus non-literal verb–noun combination [64]. For example, whereas the noun in a non-literal verb–noun combination is often non-referential, abstract, and/or predicative (as in *take time* and *give a hug*), the noun in a literal combination tends to be referential and concrete (as in *take the toys* and *give a banana*). Earlier work has used WordNet [35] to estimate non-referentiality and predicativeness by looking at the noun's position in the taxonomy, and its morphological relation to a verb [30]. However, WordNet's conceptual and lexical organization most likely does not reflect that of a child. Next, we explain two measures that instead aim to capture these properties with simple statistics over the surface behaviour of the noun.

Non-referentiality. Non-referential nouns (such as those in non-literal expressions) tend to appear in particular syntactic forms [42]—typically preceded by an indefinite determiner (such as *a/an*) or no determiner [34, 76]. Moreover, it has been shown that children indeed associate certain semantic properties with surface syntactic forms [10]. Here we assume that a noun is recognized as non-referential to the extent that it occurs in this preferred pattern of determiner use, i.e.:

[3]The choice of verb can vary among dialects of the language; for example, British speakers typically say *take a decision* instead of *make a decision* and *have a nap* instead of *take a nap*.

[4]Although it remains to be tested whether children actually do this, a construction grammar approach to language acquisition, as in Goldberg [41], supports this type of calculation, since the learner would keep track of which nouns can occur in which constructions.

$$\text{NRef}(n) \doteq P(pt_{nref} | n) = \frac{\text{freq}(n, pt_{nref})}{\text{freq}(n)} \tag{3}$$

where pt_{nref} = ⟨det:*a/an/*NULL n⟩, freq(n, pt_{nref}) is the frequency of occurrence of n in pattern pt_{nref}, and the denominator estimates the frequency of n in any pattern. Note that we look at all occurrences of a noun irrespective of its grammatical relation to a verb; this is thus a simple relative frequency for a child to determine: of the instances she sees of this noun, what proportion are in this particular pattern.

Predicativeness. In a non-literal verb–noun combination, such as *make a decision*, the predicative meaning is contributed mainly by the noun component, i.e., *decision*. Moreover, in such expressions the noun is often morphologically related to a verb (e.g., *decision* as the nominalized form of *decide*). To capture this property, previous work has looked at whether the noun has a morphologically-related verb form [30]. We cannot assume that full knowledge of morphology is in place before a child starts learning about non-literal expressions. But it has been shown that young children can accurately predict whether a word is used as a verb or a noun in a given context [10]. We thus measure predicativeness of the noun n in a verb–noun pair as the relative frequency of the form n (e.g., *push* in *give a push*) being used as a verb (as in, e.g., *push the door*).

$$\text{Pred}(n) \doteq \frac{\text{freq}(n_V)}{\text{freq}(n_V) + \text{freq}(n_N)} \tag{4}$$

where freq(n_V) is the frequency of the form n appearing as a verb, and freq(n_N) is the frequency of the form n appearing as a noun.

3.3 Degree of Syntactic Fixedness

Young children show evidence of learning associations between a complex syntactic form and a specific semantic interpretation (e.g., [36, 70]). It is thus reasonable to assume that children can use the information about the surface syntactic behaviour of a verb–noun combination to identify its semantic class. Here we devise statistical measures that aim at capturing the differing syntactic behaviour of non-literal and literal combinations.

Non-literal expressions are known to have a fixed syntactic structure and not occur in a variety of forms [20, 26]. More specifically, ABS+LVC expressions, while allowing some variation, are relatively restricted compared to LIT expressions. For example, an LVC such as *give a shout* allows limited noun and determiner variation; e.g., *give some shouts* and *give the shout* are not as acceptable as *give a shout*. This is also true for ABS expressions. For example, *take a time* and *take times* are not recognized as acceptable variations of *take time*. In contrast, literal expressions are generally much more syntactically flexible, e.g., *take an apple*, *take the apple*, and *take three apples* are all acceptable.

Although there is some variation, most LVC and ABS expressions appear in the form $pt_{fixed} = \langle v \text{ det: } a/an/\text{NULL } n \rangle$. (Note that the noun is in the same pattern as for NRef above; the difference is that here the focus is on the degree to which the particular verb–noun combination leads to the use of that pattern for the noun.) Measures of this type of syntactic fixedness have required keeping track of probability distributions over a wide range of items and patterns [6, 30]. Here, we estimate the degree of syntactic fixedness of a target verb–noun combination with a much simpler measure—the relative frequency of the pair in the preferred pattern:

$$\begin{aligned} \text{Fixed}(v,\, n) &\doteq P(pt_{fixed} |\, v,\, n,\, \text{gr} = \text{dobj}) \\ &= \frac{\text{freq}(v, n, \text{gr} = \text{dobj}, pt_{fixed})}{\text{freq}(v, n, \text{gr} = \text{dobj})} \end{aligned} \tag{5}$$

Children appear to store specific information about the frequency of occurrence of multiword sequences in general (e.g., [7]), and about verb–argument structures in particular (e.g., [41, 74]). We thus expect the above calculations to be plausible for children.

We have described five simple statistical measures that may be plausible for children to keep track of. In the remainder of the paper, we first present experiments that evaluate how well the measures can identify non-literal verb–noun combinations in child-directed speech, and then describe extensions to a word learning model that enable it to learn the meaning of such expressions by incorporating these statistical measures.

4 Evaluating the Statistical Measures

In this section, we present two types of experiments to determine the potential of our statistical measures to identify non-literal verb–noun combinations in child-directed speech. Each of our measures assigns a numerical score to the expressions that reflects one of the linguistic properties that may be useful to a child in determining which are literal and which are non-literal. To evaluate their effectiveness, we first (in Sect. 4.2) apply a hierarchical agglomerative clustering algorithm that uses the scores to separate all the experimental expressions into two clusters, and then see how closely those clusters correspond to the actual labels on the expressions as LIT, or as ABS+LVC. Since we assume that, in any learning situation, a combination of the cues might be at work, we use all five measures as input to the clustering algorithm.

The clustering results thus show the effectiveness of the measures working together to separate non-literal from literal combinations. We further analyze (in Sect. 4.3) each individual measure in its ability to separate literal and non-literal expressions, in order to better understand how relevant each measure is to the identification of multiword lexemes. We begin by presenting the details of the experimental data and evaluation methods.

4.1 Experimental Setup

Corpus. To gather input for our experiments, we use the American English section of the CHILDES database [52], removing 16 corpora that either lack child-directed speech (CDS) or belong to a special group with a particular language use (e.g., socio-economically distinguished). All the data are automatically parsed with the parser of Sagae et al. [68]. Because we are interested in what is learnable from input a child is exposed to, the statistics for all experiments are extracted from CDS. The size of the CDS portion of the corpus is about 600, 000 utterances, which contain nearly 3.2 million words (including punctuation).

Experimental expressions. In this work we focus on two basic verbs, *take* and *give*, because they are highly polysemous and frequently used in verb–noun combinations [15]. We extract verb–noun combinations that contain these verbs from the CDS portion of the data. The final expression list that is used in the experiments includes those verb–noun pairs with a frequency of at least 5. In some experiments, we further restrict the data to higher-frequency verb–noun combinations, i.e., those occurring at least 10 times. Dealing with low-frequency items is important in modeling child language acquisition, and here we vary the relatively low cutoff to see if it helps to have more items. The final list of expression types was annotated by a native English speaker with four classes: LIT, ABS, LVC, and IDM. Note that we consider expression types, not tokens. Thus, if a verb–noun combination had usages that fall into more than one class, the annotator chose the class that seemed to reflect the predominant usage.[5] Invalid expressions (due to parsing errors) and the single instance of an IDM were removed from the expression list. Table 1 presents the number of expressions in each class, as well as the total number of non-literal expressions (ABS+LVC).

Evaluation. To evaluate the clustering experiments, we assign to each resulting cluster a label (either LIT or ABS+LVC), which is the label of the majority of items in the cluster, and calculate accuracy (*Acc*) and completeness (*Comp*) as measures of the goodness of the cluster. Accuracy gives the proportion of expressions in a cluster that have the same label as the cluster; completeness gives the proportion of all expressions with the same label as the cluster that are actually placed in that cluster. (Note that *Acc* is similar to precision, and *Comp* to recall.)

Recall that our measures are designed such that each is expected to be higher for the non-literal expressions than for the literal ones. In evaluating the measures individually, we can thus use each measure to rank the expressions and see whether ABS+LVC expressions are generally ranked higher than LIT ones. We do this for

[5] For example, the verb–noun pair *give-hand* may occur as an ABS usage (*give me a hand cleaning up*) or as a LIT usage (*give me Mr. PotatoHead's hand* or *give me your pretty hands*). In most cases of such potential ambiguity, the annotator had a clear intuition of which would be the predominant usage, since the alternative would be odd to find in CDS. In some cases, such as *give-hand*, the actual corpus usages were examined to determine the most frequent class.

Table 1 A detailed breakdown of the experimental expressions

V_b	Total	LIT	ABS	LVC	ABS+LVC
198 expressions with *freq* $\geq$ 5					
take	108	77	18	13	31 (29 %)
give	90	73	7	10	17 (19 %)
take and *give*	198	150	25	23	48 (24 %)
98 expressions with *freq* $\geq$ 10					
take	57	38	8	11	19 (33 %)
give	41	30	4	7	11 (27 %)
take and *give*	98	68	12	18	30 (31 %)

take and *give* expressions separately, and for all expressions together. We use a standard evaluation metric, namely average precision (*AvgPrec*), which reflects the goodness of a measure in placing expressions from the target classes (ABS and LVC) before those from the other class (LIT), and is calculated as the average of *precision* scores at different thresholds.

We also compare the performance of each measure against a baseline which reflects how hard the task is. We randomly assign a value between 0 and 1 to each expression in a set, generating a random ranked list. We repeat this process 1,000 times and report the average of the *AvgPrec* values for each of these random lists as our baseline. We also calculate the relative error rate reduction (*ERR*) of each measure over the random baseline. To calculate *ERR* for a measure, we divide the difference between the error rates of the measure and the baseline by that of the baseline.

4.2 Measures in Combination: Clustering

Results of the clustering experiments are shown in Table 2. We can see that *Acc* for non-literal expressions is high only for the higher-frequency expressions (compare C_2 in each panel of the table). We also see that literal expressions are better separated than non-literal ones since their *Comp* score is much higher (compare C_1 and C_2 for each panel of the table). Looking closely at the number of expressions of different labels (LIT, LVC, and ABS) in each cluster, it is clear that ABS expressions are more mixed with LIT expressions compared to LVC ones. Consequently, the measures are better in separating LVC from LIT than ABS from LIT.

We performed two-way clustering on the assumption that a two-way distinction would be easier for the measures than a three-way distinction. However, the poor performance on ABS expressions may be due to a weakness of the measures, or may be due to a need for three clusters to capture the pattern in the data. We thus also performed a three-way clustering to examine the goodness of measures in dividing expressions into ABS, LVC, and LIT classes (see Table 3). According to the results, ABS expressions do not form a separate cluster, and are again mixed in

Table 2 Two-way clustering results. C_i represents Cluster i; *Label* is the majority class in the cluster; *Acc* and *Comp* are explained in the text

	LVC	ABS	LIT	*Label*	*Acc* (%)	*Comp* (%)
On 198 expressions with *freq* $\geq$ 5						
C_1	1	13	122	LIT	90	81
C_2	22	12	28	ABS+LVC	55	71
On 98 expressions with *freq* $\geq$ 10						
C_1	1	9	65	LIT	87	96
C_2	17	3	3	ABS+LVC	87	67

Table 3 Three-way clustering results. C_i represents Cluster i; *Label* is the majority class in the cluster; *Acc* and *Comp* are explained in the text

	LVC	ABS	LIT	*Label*	*Acc* (%)	*Comp* (%)
On 198 expressions with *freq* $\geq$ 5						
C_1	3	7	27	LIT	73	18
C_2	1	13	122	LIT	90	81
C_3	19	5	1	LVC	76	83

with the LIT and LVC clusters. Future work will need to verify whether this is due to an inconsistent annotation of the ABS expressions, or because our measures do not adequately capture properties of this class. Interestingly, however, a three-way clustering results in forming a more coherent LVC class: compare *Acc* and *Comp* for C_3 in Table 3 with those for C_2 in the top panel of Table 2.

4.3 Performance of the Individual Measures

We test the performance of each measure, for *take* and *give* expressions separately, and for all the expressions with *take* and *give*. The results in Table 4 show that all measures perform better than the baseline (at separating non-literal expressions from literal ones), with CProb, Pred, and Fixed having the best performance. These results suggest that simple statistical measures that draw on specific linguistic properties of non-literal verb–noun combinations—measures which are plausible for children to keep track of—can indeed be effective in recognizing non-literal expressions.

We also observe that, in general, our measures perform better on the expressions composed with *take* than the expressions with *give*. A possible explanation is that the *give* expressions are more complicated, because *give* more often occurs in a double object construction (in comparison to *take*). It remains to be tested whether children also show more difficulty in learning *give* expressions.

Looking at performance on higher-frequency expressions, we see that all measures show an improvement. However, note that only for two of the measures

Table 4 Performance (*AvgPrec*) of the individual measures. The numbers in parentheses show the *ERR* of the measures for *take* and *give* expressions combined

Measure	*take*	*give*	*take* and *give*
On 198 expressions with *freq* $\geq$ 5			
Baseline	0.28	0.19	0.24
Cooc	0.53	0.38	0.51 (0.35)
CProb	0.65	0.47	0.56 (0.42)
NRef	0.49	0.33	0.40 (0.21)
Pred	0.59	0.54	0.59 (0.46)
Fixed	0.66	0.44	0.56 (0.40)
On 98 expressions with *freq* $\geq$ 10			
Baseline	0.33	0.27	0.31
Cooc	0.57	0.41	0.54 (0.34)
CProb	0.71	0.57	0.64 (0.48)
NRef	0.62	0.49	0.55 (0.35)
Pred	0.68	0.59	0.67 (0.52)
Fixed	0.84	0.56	0.71 (0.58)

(NRef and Fixed) the gain in performance is substantially more than the increase in the baseline performance. These two measures summarize the syntactic behaviour of a word or a combination by examining all their usages. For higher-frequency expressions (with more usages), it is possible that the evidence available for these measures is more reliable, resulting in better performance.

5 Embedding the Measures into a Word Learning Model

The results presented so far suggest that simple statistics over the usages of a verb–noun combination (and its components) have the potential to provide useful cues for a child to identify non-literal expressions. We need to explore further how children learning the vocabulary of their native language might use such statistical cues to recognize that certain combinations of words in their input actually form multiword lexemes. We investigate this issue by incorporating (some of) the statistical measures into the operations of an existing computational model of early word learning in children, namely, that of [33].

We first give a brief overview of the original word learning model in Sect. 5.1 (we refer the interested reader to [33] for a full explanation of this model). When processing a multiword lexeme, such as *take a nap*, the original model finds a meaning for each individual word (*take*, *a*, *nap*) just as it does for a literal combination of words, such as *take any toys*. There is no mechanism for the model

to associate a single meaning with the sequence of words *take a nap*.[6] We thus add a preprocessing step, described in Sect. 5.2, in which the model draws on statistics collected thus far to decide whether a given sequence of words in the input utterance should be considered as a multiword lexeme. Section 5.3 presents an evaluation of the new model with respect to the acquisition of multiword lexemes of the form verb–noun.

5.1 The Original Word Learning Model

We use the model of Fazly et al. [33], which is a probabilistic incremental model of cross-situational word learning in children. The input to the model is a list of pairs of an utterance (what the child hears, represented as a set of words) and a scene (what the child perceives or conceptualizes, represented as a set of meaning symbols), as in:[7]

Utterance: *Joe is happily eating an apple*
Scene: JOE, IS, HAPPILY, EAT, A, APPLE

The model incrementally learns a meaning for each word in the input as a probability distribution over all meaning symbols, $P(m|w)$, referred to as the *meaning probability* of the word, as in:

Prior to receiving any usages of a given word, the model assumes that all symbols have equal probability as its meaning. The model then updates the meanings of words by processing each utterance–scene pair in two steps.

As the first step in processing an input utterance–scene pair, the model, like children, must determine which meaning symbol in the scene is associated with

[6]The original model of Fazly et al. treats utterances as unordered bags of words, ignoring syntactic information. Syntax is arguably a valuable source of knowledge in word learning in children (e.g., [39,56]). In a preliminary study, Alishahi and Fazly [2] also show that the word learning model can potentially benefit from knowledge of syntactic categories. Such information might be necessary for the acquisition of multiword lexemes, and should be further investigated in the future.

[7]Following Fazly et al. [33] we assume that words such as *a* and *is* also have corresponding meaning symbols in the scene. Such words are often considered by linguists to mainly have a grammatical function. However, it is reasonable to assume that language learners perceive some aspects of their meaning (e.g., definite/indefinite for a determiner such as *a*, and state/action for the verb *be*) from the scene.

each word in the utterance. (Note that the input does not indicate which meaning goes with which word.) This process is called the *alignment* of words and meaning symbols. Alignment is probabilistic, so that each word is aligned more or less strongly with each meaning, according to the model's partially-learned knowledge of meaning probabilities as calculated thus far. Specifically, the probability of aligning a meaning symbol and a word in the current input is proportional to the current meaning probability of that meaning symbol for the word, and is disproportional to the meaning probabilities of the meaning symbol and the other words in the utterance. That is, a word w and meaning m are strongly aligned if $P(m|w)$ is relatively high and $P(m|w')$ is relatively low for other words w' in the utterance.

As the second step, the meaning probabilities of the words in the current utterance are updated according to the accumulated (probabilistic) evidence from prior co-occurrences of words and meaning symbols (reflected in the alignment probabilities). This evidence is collected by maintaining a running total of the alignment probabilities over all input pairs encountered so far, yielding an accumulated frequency of co-occurrence of a word–meaning pair, weighted by the strength of alignment between the two each time they are observed together. Meaning probabilities for current words are then re-calculated from these incrementally-accumulated alignment probabilities.

5.2 Learning the Verb–Noun Multiword Lexemes

The approach described above learns a separate meaning probability distribution for each word. To enable the model to learn a meaning distribution for a verb–noun combination such as *give a kiss*, the model must be able to identify the expression as a single unit of meaning. To achieve this, we add an input pre-processing step to the original model and slightly modify the way alignment probabilities are calculated.

We assume that upon receiving an utterance–scene pair containing any verb–noun combination (literal or non-literal), a learner (here the model) simultaneously considers two possible interpretations: That the verb–noun combination is a multiword lexeme, or that the combination is literal. That is, when the original model receives an input such as:

U : *give me a kiss*
S : GIVE, ME, A, KISS

our modified model will also consider the alternative interpretation in which the verb and noun form a single unit of meaning:

U′ : *give-kiss me a*
S′ : ME, A, GIVE-KISS

This alternative interpretation is created by merging the verb and the noun into a single word (*give-kiss*), and by creating a new meaning symbol for the associated

event (GIVE-KISS). We assume that the learner has a certain confidence in either of these interpretations given what has been learned about words and meanings in the input thus far. Specifically, the learner calculates a probability $\text{prob}_{\text{mwl}}(v, n)$ which reflects its confidence that the verb–noun combination in the utterance is a non-literal multiword lexeme, as in (U′–S′) above. This probability combines the two statistical measures, namely CProb and Pred, which were the best in separating literal and non-literal expressions in our earlier experiments.[8] More formally, $\text{prob}_{\text{mwl}}(v, n)$ is computed as in:

$$\text{prob}_{\text{mwl}}(v, n) = \alpha * \text{CProb}(v, n) + (1 - \alpha) * \text{Pred}(n)$$

where α is set to 0.5, weighting the evidence from the two statistical measures equally. Thus, the interpretation that a verb–noun combination is a multiword lexeme, as in (U′–S′) above, is assigned a confidence score equal to $\text{prob}_{\text{mwl}}(v, n)$, and the other interpretation, as in (U–S) above, is given the confidence score of $1 - \text{prob}_{\text{mwl}}(v, n)$.

Whenever there is a verb–noun pair in an utterance, we calculate separate alignment probabilities over two possible utterance–scene pairs corresponding to the two interpretations. The two sets of alignment probabilities are then combined, using $\text{prob}_{\text{mwl}}(v, n)$ as a weight, to get a single alignment probability for each word and meaning symbol in the input pair:

$$\begin{aligned}\text{align}(w|m) = {} & \text{prob}_{\text{mwl}}(v, n) * \text{align}_1(w|m) \\ & + (1 - \text{prob}_{\text{mwl}}(v, n)) * \text{align}_2(w|m)\end{aligned}$$

Note that for a *w–m* pair that occurs only in one interpretation (e.g., *give-kiss*–GIVE-KISS), its alignment would be zero in the other interpretation. This means that the learner aligns each word and meaning symbol to the extent that it is confident that the corresponding interpretation is accurate. The modified alignment probabilities are then used to calculate the meaning probabilities as in the original model.

5.3 Experiments on the Modified Word Learner

We expect the modified word learning model to learn a single meaning for non-literal verb–noun pairs but not for literal ones. That is, we expect a meaning probability such as P(GIVE-KISS|*give-kiss*) to be high, since *give-kiss* is a multiword lexeme that expresses a kissing event. By contrast, P(GIVE-PRESENT|*give-present*)

[8] We did not incorporate the Fixed measure into this probability, because this measure needs to consider the usage pattern across several occurrences, and many of the experimental items in this corpus have frequency of only 1 or 2.

Table 5 The number and percentage of verb–noun combinations in each class that are learned correctly: i.e., as literal for the LIT class, and as non-literal for the ABS and LVC classes

Class	Size	Learned correctly	
		Number	Percentage (%)
LIT	115	105	91
ABS	24	8	33
LVC	32	24	75

should be low, since *give a present* is literal with individual associations of give to GIVE and present to PRESENT.

We use the same data as in Fazly et al. [33]: 180, 499 utterance–scene pairs, where the utterances are taken from the Manchester corpus in the CHILDES database [52], and the scene representations are automatically constructed using an input-generation lexicon containing a symbol as the meaning of each word. Because the Manchester corpus is British English and some American English verb–noun multiword lexemes with *take* occur with other basic verbs in British English, we only consider the verb–noun combinations with *give* in the current experiments. Since children can learn meanings of very low frequency words, we do not apply a frequency cut-off, but rather consider all verb–noun combinations with *give* in the corpus. The number of LIT, ABS, and LVC expressions used in our experiments is shown in Table 5.

In Fazly et al. [33], a word–meaning pair is considered learned if the probability of the correct meaning given the word is above 0.7. This is a somewhat arbitrary cut-off, but to be consistent we use the same threshold. We say that a verb–noun combination with *verb* and *noun* is "learned as a multiword lexeme" if the probability P(VERB-NOUN|*verb-noun*) is above this threshold—that is, the combination of the verb and noun words are associated with a single (correct) meaning. We say that a verb–noun combination is "learned correctly" if the combination is non-literal and is learned as a multiword lexeme, or the combination is literal and is **not** learned as a multiword lexeme. To evaluate the model's ability in learning multiword lexemes, we look at the proportion of expressions from each class that are learned correctly; see Table 5.

The results in Table 5 show that the model performs very well on the LVC and LIT expressions (75 % and 91 %, respectively), but only a small proportion (33 %) of the ABS expressions are learned correctly. A closer look at the results shows that many of the non-literal expressions with a low frequency of 1 are not learned correctly. This includes 46 % of LVC expressions with frequency 1, and 85 % of ABS expressions with frequency 1. This finding is in line with what has been observed in children: that children are faster at producing more familiar (frequent) multiword sequences [7]. It remains to be tested whether children also are unable to learn some of these MWEs (as MWEs) from a single exposure.

6 Conclusions

Our results confirm that simple statistical measures that draw on linguistic properties of non-literal expressions are useful in identifying them. The best measure for *give* and *take* expressions is Pred, i.e., the normalized frequency of the usages of the noun as a verb. The success of this measure indicates that the predicativeness of the noun is a salient property of non-literal verb–noun combinations. The goodness of CProb in identifying non-literal expressions suggests that the verb–noun pair in such expressions is more entrenched compared to literal ones and exhibits collocational behaviour. However, collocational behaviour alone is not a very good indicator of non-literal expressions; the CProb measure consistently outperforms Cooc (which only quantifies the entrenchment of the verb–noun pair). The key difference between these two measures is that in CProb, we also measure the degree that the noun selects for the appropriate verb. The Fixed measure which looks at a specific syntactic pattern for non-literal expressions performs as well as CProb for all expressions, but is the best measure for expressions having frequency of at least ten, for which there is sufficient evidence of typical syntactic usage.

Our measures are generally better for higher-frequency expressions. However, two of the best measures (Pred and CProb) perform well on both expressions with frequency of at least 5 and higher-frequency expressions, suggesting that children might be able to learn verb–noun combinations even with very little input. Our results also show that the performance of our measures is better for *take* expressions compared to *give*. The Fixed measure especially performs well on *take*, but less well on *give*, suggesting that the more complex syntactic constructions that *give* appears in (e.g., the double object construction) may cause children difficulty.

We also integrate our measures into a word learning model, and show that the new model can successfully learn the meaning of many LVC expressions. Future work will need to further investigate why it is harder for the model to learn the meaning of ABS expressions. In the experiments presented in this article, we have focused on a small number of verb–noun combinations (namely, 117) formed around one particular verb (i.e., *give*). To better understand the generalizability of our findings, future research will need to extend these experiments to other verbs (e.g., *take*) and to other types of multiword lexemes (e.g., noun compounds).

Another limitation of the model is that it learns word meanings by mapping each word to a distinct 'concept' (e.g., *give-kiss* must be mapped to GIVE-KISS). In the future, we need to use a richer semantic representation where each concept is comprised of finer-grained semantic primitives. The use of such a representation would enable the model to determine semantic similarities among words (e.g., the similarity between the meaning of the expression *give-kiss* and that of the verb *kiss*), which would further allow it to make generalizations across different types of lexical items.

References

1. Alba-Salas, J. (2002). *Light verb constructions in Romance: A syntactic analysis*. Ph.D. thesis, Cornell University.
2. Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In *Proceedings of CogSci'2010*, Portland.
3. Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science: A Multidisciplinary Journal, 32*(5), 789–834.
4. Alishahi, A., & Stevenson, S. (2011). Gradual acquisition of verb selectional preferences in a Bayesian model. In Poibeau et al. (2011).
5. Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82. ISSN 0749–596X.
6. Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Multiword Expression'07: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions* (pp. 1–8). Prague: Association for Computational Linguistics.
7. Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*(3), 241–248.
8. Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 65–72), Sapporo.
9. Borensztajn, G., Zuidema, W., & Bod, R. (2009). Children's grammars grow more abstract with age – evidence from an automatic procedure for identifying the productive units of language. *Topics in Cognitive Science, 1*(1), 175–188.
10. Brown, R. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal Psychology, 55*(1), 1–5.
11. Brown, R. (1973). *A first language: The early stages*. Cambridge: Harvard University Press.
12. Butt, M. (1997). Aspectual complex predicates, passives and dispositionability. In *Talk Held at the 1997 Meeting of the Linguistics Association of Great Britain (LAGB'97)*, University of Essex. http://ling.uni-konstanz.de/pages/home/butt/.
13. Chang, N. (2004). Putting meaning into grammar learning. In *Proceedings of the ACL'04 Workshop on Psycho-Computational Models of Human Language Acquisition* (pp. 17–24), Geneva.
14. Church, K., Gale, W., Hanks, P., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Exploiting on-line resources to build a lexicon* (pp. 115–164). Hillsdale: Erlbaum.
15. Claridge, C. (2000). *Multiword verbs in early modern english. Language and Computers 32*. New York: Rodopi.
16. Clark, E. V. (1996). Early verbs, event-types, and inflections. In C. E. Johnson & J. H. V. Gilbert (Eds.), *Children's language* (Vol. 9, pp. 61–73). Mahwah: Erlbaum.
17. Clark, A. (2001). Unsupervised induction of stochastic context free grammars with distributional clustering. In *Proceedings of Conference on Computational Natural Language Learning* (pp. 105–112), Toulouse.
18. Connor, M., Fisher, C., & Roth, D. (2011). Starting from scratch in semantic role labeling: Early indirect supervision. In Poibeau et al. (2011).
19. Cook, P., & Stevenson, S. (2006). Classifying particle semantics in English verb-particle constructions. In *Proceedings of the COLING-ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties* (pp. 45–53), Sydney.
20. Cowie, A. P. (1981). The treatment of collocations and idioms in learner's dictionaries. *Applied Linguistics, II*(3), 223–235.
21. Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 605–613), Ann Arbor.

22. Devereux, B. J. & Costello, F. J. (2011). Learning to interpret novel noun-noun compounds: Evidence from category learning experiments. In Poibeau et al. (2011).
23. Dominey, P. F., & Inui, T. (2004). A developmental model of syntax acquisition in the construction grammar framework with cross-linguistic validation in English and Japanese. In *Proceedings of the ACL'04 Workshop on Psycho-Computational Models of Human Language Acquisition* (pp. 33–40), Geneva.
24. Dras, M. (1995). Automatic identification of support verbs: A step towards a definition of semantic weight. In *Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence* (pp. 451–458). Singapore: World Scientific.
25. Dras, M., & Johnson, M. (1996). Death and lightness: Using a demographic model to find support verbs. In *Proceedings of the Fifth International Conference on the Cognitive Science of Natural Language Processing* (pp. 165–172), Dublin.
26. Everaert, M., van der Linden, E. -J., Schenk, A., & Schreuder, R. (Eds.). (1995). *Idioms: Structural and psychological perspectives*. Hillsdale: Lawrence Erlbaum Associates.
27. Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook*. Berlin: Mouton de Gruyter. Article 58.
28. Evert, S., Heid, U., & Spranger, K. (2004). Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th Int'l Conference on Language Resources and Evaluation* (pp. 907–910), Lisbon.
29. A. Fazly. (2007). *Automatic acquisition of lexical knowledge about multiword predicates*. Ph.D. in Computer Science, University of Toronto.
30. Fazly, A., & Stevenson, S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Multiword Expression'07: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions* (pp. 9–16), Prague. Association for Computational Linguistics.
31. Fazly, A., Stevenson, S., & North, R. (2007). Automatically learning semantic knowledge about multiword predicates. *Journal of Language Resources and Evaluation, 41*(1), 61–89.
32. Fazly, A., Nematzadeh, A., & Stevenson, S. (2009). Acquiring multiword verbs: The role of statistical evidence. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam.
33. Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science, 34*, 1017–1063.
34. Fellbaum, C. (1993). *The determiner in English idioms* (pp. 271–295). Hillsdale: Lawrence Erlbaum Associates.
35. Fellbaum, C. (Ed.). (1998). *WordNet, an electronic lexical database*. Cambridge/London: MIT Press.
36. Fisher, C. (2002). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science, 5*(1), 55–64.
37. Frank, M., Goodman, N., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*. Cambridge/London: MIT
38. Gentner, D., & France, I. M. (2004). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M. K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence* (pp. 343–382). San Mateo: Kaufmann.
39. Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*(8), 684–691.
40. Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
41. Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
42. Grant, L. E. (2005). Frequency of 'core idioms' in the British National Corpus (BNC). *International Journal of Corpus Linguistics, 10*(4), 429–451.

43. Grefenstette, G., & Teufel, S. (1995). Corpus-based method for automatic identification of support verbs for nominalization. In *Proceedings of the 7th Meeting of the European Chapter of the Association for Computational Linguistics (EACL'95)* (pp. 98–103), Dublin.
44. Israel, M. How children get constructions. In M. Fried & J. -O. Ostman (Eds.), *Pragmatics in construction grammar and frame semantics*. John Benjamins. (submitted)
45. Karimi, S. (1997). Persian complex verbs: Idiomatic or compositional? *Lexicology, 3*(1), 273–318.
46. Kearns, K. (2002). Light verbs in English. unpublished manuscript. http://www.ling.canterbury.ac.nz/people/kearns.html.
47. Krott, A., Gagne, C., & Nicoladis, E. (2009). How the parts relate to the whole: Frequency effects on childrens interpretations of novel compounds. *Journal of Child Language, 36*(01), 85–112.
48. Kytö, M. (1999). *Collocational and idiomatic aspects of verbs in Early Modern English* (pp. 167–206). Amsterdam/Philadelphia: John Benjamins Publishing Company.
49. Xiaowei, P. Li, & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science, 31*, 581–612.
50. Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 317–324), College Park. Association for Computational Linguistics.
51. Lin, T. -H. (2001). *Light verb syntax and the theory of phrase structure*. Ph.D. thesis, University of California, Irvine.
52. MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. The Database* (3rd ed., Vol. 2). Mahwah: Lawrence Erlbaum Associates.
53. McCarthy, D., Keller, B., & Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 73–80), Sapporo.
54. Miyamoto, T. (2000). *The light verb construction in Japanese: The role of the verbal noun*. Amsterdam/Philadelphia: John Benjamins.
55. Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. New York: Oxford University Press.
56. Naigles, L., & Kako, E. T. (1993). First contact in verb acquisition: Defining a role for syntax. *Child Development, 64*, 1665–1687.
57. Nation, K., Marshall, C. M., & Altmann, G. T. M. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology, 86*, 314–329.
58. Newman, J. (1996). *Give: A cognitive linguistic study*. Berlin/New York: Mouton de Gruyter.
59. Newman, J., & Rice, S. (2004). Patterns of usage for English SIT, STAND, and LIE: A cognitively inspired exploration in corpus linguistics. *Cognitive Linguistics, 15*(3), 351–396.
60. Onnis, L., Roberts, M., & Chater, N. (2002). Simplicity: A cure for overgeneralizations in language acquisition. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 720–725), Fairfax.
61. Parisien, C., & Stevenson, S. (2010). Learning verb alternations in a usage-based Bayesian model. In *Proceeding of the 32nd Annual Meeting of the Cognitive Science Society*, Austin.
62. Pauwels, P. (2000). *Put, set, lay and place: A cognitive linguistic approach to verbal meaning*. Munich: Lincom Europa.
63. Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37*(3), 607–642.
64. Quochi, V. (2007). *A usage-based approach to light verb constructions in Italian: Development and use*. Ph.D. thesis, Universit'a di Pisa.
65. Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science, 29*, 819–865.
66. Riehemann, S. (2001). *A constructional approach to idioms and word formation*. Ph.D. thesis, Stanford University, Stanford.

67. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)* (pp. 1–15), Mexico City, Mexico.
68. Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proceedings of the ACL'07 Workshop on Cognitive Aspects of Computational Language Acquisition*, Prague.
69. Sakas, W., & Fodor, J. D. (2001). The structural triggers learner. In S. Bertolo (Eds.), *Language acquistion and learnability*, (172–233). Cambridge: Cambridge University Press.
70. Scott, R. M., & Fisher, C. (2009). Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and Cognitive Processes, 24*, 777–803
71. Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*(1), 143–177.
72. Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language, 83*, 227–236.
73. Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2002). Going, going, gone: The acquisition of the verb 'go'. *Journal of Child Language, 29*, 783–811.
74. Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.
75. Venkatapathy, S., & Joshi, A. (2005). Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceeding of HLT-EMNLP'05* (pp. 899–906), Vancouver.
76. Wierzbicka, A. (1982). Why can you Have a Drink when you can't *Have an Eat? *Language, 58*(4), 753–799.
77. Yu, C., & Smith, L. B. (2006). Statistical cross-situational learning to build word-to-world mappings. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver.

Starting from Scratch in Semantic Role Labeling: Early Indirect Supervision

Michael Connor, Cynthia Fisher, and Dan Roth

Abstract A fundamental step in sentence comprehension involves assigning semantic roles to sentence constituents. To accomplish this, the listener must parse the sentence, find constituents that are candidate arguments, and assign semantic roles to those constituents. Where do children learning their first languages begin in solving this problem? To experiment with different representations that children may use to begin understanding language, we have built a computational model for this early point in language acquisition. This system, Latent BabySRL, learns from transcriptions of natural child-directed speech and makes use of psycholinguistically plausible background knowledge and realistically noisy semantic feedback to improve both an intermediate syntactic representation and its final semantic role classification. Using this system we show that it is possible for a simple learner in a plausible (noisy) setup to begin comprehending the meanings of simple sentences, when initialized with a small amount of concrete noun knowledge and some simple syntax-semantics mapping biases, before acquiring any specific verb knowledge.

1 Introduction

When learning their first language, children must cope with enormous ambiguity in both the meaning and structure of input sentences. Ultimately, children must select candidate meanings by observing the world and align them with the sentence presented in the input. They must do so without already knowing which parts of the

M. Connor (✉) · D. Roth
Department of Computer Science, University of Illinois, Urbana, USA
e-mail: connor2@illinois.edu; danr@illinois.edu

C. Fisher
Department of Psychology, University of Illinois, Champaign, USA
e-mail: clfishe@illinois.edu

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_10,

sentence refer to which parts of their conceptual representations of world events. Even worse, the child must also identify the ways in which structural aspects of sentences, which are not clearly displayed in the surface form of the utterance, convey aspects of the relational meanings of those sentences. For example, phrase order or case marking identify the roles that particular constituents play in the sentence's meaning, thus conveying who does what to whom. Despite both of these sources of ambiguity, semantic and syntactic, children do learn to interpret sentences, and do so without detailed feedback about whether their interpretations, or their hypothesized syntactic structures, were correct. When faced with an ambiguous world, and with word-strings rather than sentence structures, how can learners begin to identify and interpret the syntactic structures of sentences?

The ambiguity of word-strings as evidence for syntax is a nearly universally recognized problem for language acquisition. But the ambiguity of scenes as evidence for sentence meaning is sometimes overlooked. To illustrate, take the sentence "The girl tickled the boy," accompanied by a scene in which a boy and girl play together, and at some point the girl does tickle the boy. Any scene offers up a host of candidate interpretations, both related and unrelated to the target event described by the sentence. These might include the boy and girl playing, the boy squirming and giggling, the girl giggling, background facts about the boy or girl that might be of interest (e.g., "You know that girl from preschool."), and so forth. Among the available construals might be some that are very difficult to tease apart based on even an extended sequence of suitable scenes. For example, scenes of 'giving' nearly always also involve 'getting', scenes of 'chasing' involve 'fleeing,' and scenes of 'putting' an object in a location also imply that the object 'goes' into that location. The basic predicate-argument semantics of a sentence are not simple descriptions of scenes, but rather express the speaker's selected perspective on that scene (e.g., [19]). It is up to the speaker to direct the attention of the child listener to the correct interpretation, through various means such as looking, gestures [66] and *the sentence itself*.

In this chapter we develop a computational language learner that must cope with this ambiguity of both scene and sentence in learning to classify abstract semantic roles for verbal predicate arguments. This computational learner, our 'Latent BabySRL', learns from child directed speech transcripts and ambiguous semantic feedback, treating an intermediate syntactic representation as a latent structure that must be learned along with the semantic predictions. This system allows us to test various plausible sources of knowledge and representation for the child learner, showing that simple structural cues regarding the identification of nouns are necessary for disambiguating noisy semantics.

1.1 Addressing the Ambiguity of Sentences and Scenes: Semantic and Syntactic Bootstrapping

A vivid illustration of the ambiguity of scenes comes from 'human simulation' experiments devised by Gleitman and colleagues to investigate word learning based

on observations of accompanying scenes [40, 80]. In these experiments, adult observers watched video-clips of mothers interacting with toddlers. In each video-clip, the mother had uttered a common noun or verb; the soundtracks of the videos were removed and participants heard only a 'beep' at the point in each video when the target word had been spoken. The observer's task was to guess what word the mother had said. Observers saw a series of such clips for each target word; thus they had opportunities for cross-situational observation. Participants were much more accurate in guessing the target nouns than the verbs. Performance with verbs improved considerably, however, when participants also received information about the sentence structures in which the verbs occurred. These results suggest that scene observations are systematically less informative for learning verbs than for learning nouns. The referents of many concrete nouns can be identified via scene observation alone, but verb referents are typically more abstract (i.e. less 'imageable') [40], and therefore naturally harder to observe in scenes. Efficient verb learning depends on support from sentence-structure cues.

On the other hand, despite the ambiguity of scenes, it is clear that a substantial part of the evidence required to learn a language must come from observing events. Only by observing words used in appropriate referential contexts (e.g., 'feed' when feeding is relevant, 'cookie' when cookies are relevant) could children attach appropriate semantic content to those words. For this reason, theories of language acquisition routinely assume that learning to use words and syntax in sentence interpretation is a partly supervised task, where the supervision comes from observation of world events: For each input sentence, learners use their existing knowledge of words and sentence structure to generate a possible meaning; the fit of this meaning with the referential context provides feedback for improving the child's lexical and grammatical knowledge. We would argue, however, that most theories or models of language acquisition finesse the true ambiguity of observation of world events, by assuming that the child has access to the correct interpretation of input sentences some substantial proportion of the time – often enough to support robust acquisition (e.g.,[14, 70, 83]).

The semantic bootstrapping theory is a special case of such accounts [70, 71]. The semantic bootstrapping theory focuses on the *ambiguity of word-strings* as evidence for syntactic structure, and proposes that learners are equipped with innate links between semantic and syntactic categories and structures; these links allow them to use semantic evidence to identify words and structures that are of particular syntactic types in their native language. To take a simple example, children might infer that words referring to entities in the world are nouns, or that a phrase referring to an agent of action in the main event described by an input sentence is the grammatical subject. On this account, access to word and sentence meaning (derived from scene observations) plays a privileged role in identifying syntactic structure, with the aid of innate links between syntax and semantics. To return to our 'tickle' example, the child would use previously acquired knowledge of the content words in this sentence ('girl', 'boy', and 'tickle') to choose the relevant construal of the scene. Via semantic bootstrapping, the child would then infer that the noun-phrase naming the agent of tickling should be the grammatical subject of the sentence.

The sentence "The girl tickled the boy" would then yield a data point that the child could begin to use to determine where to find the grammatical subject in English sentences.

The syntactic bootstrapping theory [54,65], in contrast, focuses on the *ambiguity of scenes*, particularly with regard to learning the abstract relational meanings of verbs and of sentence-structural devices such as word order or case marking [40]. Syntactic bootstrapping proposes that children use partial knowledge of sentence structure to select likely meanings of input sentences; by doing so they gain access to syntactic support for verb learning. Like semantic bootstrapping, syntactic bootstrapping requires that the learner have access to links between syntax and semantics; for syntactic bootstrapping to play a role in the initial creation of a lexicon and grammar, some of these links must be innate. The nature of these links is typically assumed to follow from the fundamental nature of the relational meanings of verbs [34, 40, 54, 65]: Verbs are argument-taking predicates, and the number of semantic arguments required to play out the meaning of each verb is systematically related to the phrasal structure of sentences containing that verb (e.g., [55, 71]). In our 'tickle' example, the presence of two noun-phrase arguments in the target sentence "The girl tickled the boy" is clearly no accident, but reflects the underlying predicate-argument structure of the verb.

1.2 How Could Syntactic Bootstrapping Begin?

But given the dual problem we started with, the rampant ambiguity of both word-strings and scenes, how could any aspects of sentence structure begin to guide sentence interpretation without considerable prior learning about the syntax and morphology of the native language? The 'structure-mapping' account of the origins of syntactic bootstrapping [35] proposes one way in which sentence structures might first guide sentence interpretation, even before children learn much about the syntax of the native language.

First, the structure-mapping account proposes that children are predisposed to align each noun in a sentence with a core semantic argument of a predicate. Given this bias, the number of nouns in the sentence becomes intrinsically meaningful to toddlers. In our 'tickle' illustration, simply identifying the target sentence as containing two nouns should prompt children to select an interpretation with two core participant roles. This simple constraint allows a skeletal representation of sentence structure, grounded in the learning of some nouns, to guide sentence interpretation essentially from the start – and to do so without requiring prior knowledge of verb meanings. This simple inference would yield a probabilistic distinction between transitive and intransitive sentences, increasing the probability that children interpret an input sentence as its speaker intended, despite the ambiguity of scenes. In turn, this increased accuracy in sentence interpretation puts the child in a better position to obtain useful information from the observed scene about other aspects of the meaning of the sentence. Such experiences provide useful

information about 'tickle,' and about the interpretation of English sentences more generally.

Second, the structure-mapping account, like any form of syntactic bootstrapping, assumes that children represent their experience with language in usefully abstract terms. These abstract representations both give children access to the proposed innate bias to align nouns with participant-roles [88], and permit rapid generalization of language-specific learning to new sentences and new verbs [38, 70]. As a result, each advance in learning the syntactic choices of the native language offers new constraints on verb and sentence interpretation. The structure-mapping account proposes that even skeletal representations of sentence structure grounded in a set of nouns provide a preliminary format for further learning about the syntax of the native language (see also [4]). To illustrate, experiences like the one sketched in our 'tickle' example, given abstract representations of both (partial) sentence structure and semantic roles, could provide the learner with evidence that the first of two noun arguments is an agent of action, and the second is a patient or recipient of action.

This process exemplifies the kind of iterative, opportunistic learning from partial knowledge that inspired the term 'bootstrapping'. It naturally incorporates aspects of both semantic and syntactic bootstrapping [40]: Children are assumed to identify the referents of some concrete nouns via a word-to-world mapping unaided by syntactic bootstrapping. As a result, early vocabularies tend to be dominated by nouns [37]. Children then assume, by virtue of the referential meanings of these nouns, that the nouns are candidate arguments of verbs. This is a simple form of semantic bootstrapping, requiring the use of built-in assumptions about syntax-semantics links to identify the grammatical function of known words – nouns in particular [70]. In this way, an initial noun vocabulary grounds a preliminary estimate of the syntax of the sentence, which in turn permits further word and syntax learning, via syntactic bootstrapping.

In this chapter we use a Semantic Role Labeling (SRL) task [12] based on child-directed speech (CDS) to model these initial steps in syntactic bootstrapping. Computational models of semantic role labeling face a learning problem similar to the one children face in early sentence comprehension: The system learns to identify, for each verb in a sentence, all constituents that fill a semantic role, and to determine their roles, such as agent, patient or goal. Our 'BabySRL' system [21–23] learns to predict the semantic roles of verbs' arguments in input sentences by directly implementing the assumptions of the 'structure-mapping' account. That is, the model (1) assumes that each noun is a candidate argument of a verb and (2) models the semantic prediction task with abstract role labels and abstract (though partial) sentence-representations grounded in a set of nouns. Our goals in implementing these assumptions in a computational model of semantic role labeling were to test the main claims of our account by explicitly modeling learning based on the proposed skeletal description of sentence structure, given natural corpora of child-directed speech. We equipped the model with an unlearned bias to map each noun onto an abstract semantic role, and asked whether partial representations grounded in a set of nouns are useful as a starting point in learning to interpret sentences. We used English word-order as a first case study: Can the BabySRL learn useful facts

about English sentence-interpretation, such as that the first of two nouns tends to be an agent? Crucially, in the present modeling experiments we asked whether learning that begins with the proposed representational assumptions can be used to improve the skeletal sentence representations with which the learner began.

In carrying out the simulations described here, our main preoccupation has been to find ways for our model to reflect both the ambiguity of scene-derived feedback about the meaning of input sentences and the ambiguity of word-strings as evidence for syntactic structure. Like the major theoretical accounts of language acquisition briefly discussed above, computational language-learning systems (including both those in the Natural Language Processing (NLP) tradition and more explicitly psycholinguistically-inspired models) often rely on implausibly veridical feedback to learn, both in divining syntactic structure from a sentence and in fitting a meaning to it. For example, the state-of-the-art SRL system which we used as a baseline for designing our BabySRL [72], like other similar systems, models semantic-role labeling in a pipeline model, involving first training a syntactic parser, then training a classifier that learns to identify constituents that are candidate arguments based both on the output of the preceding syntactic parser and on direct feedback regarding the identity of syntactic arguments and predicates. Features derived from the output of this closely supervised syntactic predicate-argument classifier then serve as input to a separate semantic-role classifier that learns to assign semantic roles to arguments relative to each predicate, given feedback about the accuracy of the role assignments. At each level in this traditional pipeline architecture, the structure that is learned is not tailored for the final semantic task of predicting semantic roles, and the learning depends on the provision of detailed feedback about both syntax and semantics. In essence, whereas children learn through applying partial knowledge at multiple levels of the complex learning and inference problem, successful computational learners typically require incorporating detailed feedback at every step. Therefore our first steps in developing the BabySRL have been to simplify the representations and the feedback available at each step, constrained by what we argue is available to children at early points in language learning (see below).

In the present work we built a computational system that treats a simple form of syntax as a hidden structure that must be learned jointly with semantic role classification. Both types of learning are based on the representational assumptions of the structure-mapping account, and on the provision of high-level, but varyingly ambiguous, semantic feedback. To better match the learning and memory capabilities of a child learner, we implemented our learning in an online, sentence-by-sentence fashion.

With this system we aim to show that:

- Nouns are relatively easy to identify in the input, using distributional clustering and minimal supervision.
- Once some nouns are identified as such, those nouns can be used to identify verbs based on the verbs' argument-taking behavior.
- The identification of nouns and verbs yields a simple linear sentence structure that allows semantic-role predictions.

- The skeletal sentence structure created via minimally supervised noun identification provides constraints on possible sentence structures, permitting the Latent BabySRL to begin learning from highly ambiguous semantic-role feedback.

2 BabySRL and Related Computational Models

In our previous computational experiments with the BabySRL, we showed that it is possible to learn to assign abstract semantic roles based on shallow sentence representations that depend only on knowing the number and order of nouns; the position of the verb, once identified, added further information [21]. Table 1 gives an example of the representations and feedback that were originally used to drive learning in the BabySRL. In our first simulations [21], full (gold standard) semantic-role feedback was provided along with a shallow syntactic input representation in which nouns and verbs were accurately identified. This skeletal representation sufficed to train a simple semantic role classifer, given samples of child-directed speech. For example, the BabySRL succeeded in interpreting transitive sentences with untrained (invented) verbs, assigning an agent's role to the first noun and a patient's role to the second noun in test sentences such as "Adam krads Mommy". These first simulations showed that representations of sentence structure as simple as 'the first of two nouns' are useful as a starting point for sentence understanding, amid the variability of natural corpora of child-directed speech.

However, the representations shown in Table 1 do not do justice to the two sources of ambiguity that face the human learner, as discussed above. The original BabySRL modeled a learner that already (somehow) knew which words were nouns and in some versions which were verbs, and also could routinely glean the true interpretation of input sentences from assumed observation of world events. These are the kinds of input representations that make syntactic and semantic bootstrapping unnecessary (in the model), and that we have argued are not available to the novice learner. Therefore in subsequent work, we began to weaken these assumptions, reducing the amount of previous knowledge assumed by the input representations and by the semantic-role feedback provided to the BabySRL. These next steps showed that the proposed simple structural representations were robust to drastic reductions in the integrity of the semantic-role feedback (when gold-standard semantic role feedback was replaced with a simple animacy heuristic for identifying likely agents and non-agents; [22]) or of the system for argument and predicate identification (when gold standard part-of-speech tagging was replaced with a minimimally-supervised distributional clustering procedure; [23]). In this chapter we develop a system that learns the same semantic role labeling task when given input representations and feedback that in our view more closely approximate the real state of the human learner: semantic feedback that is dramatically more ambiguous, coupled with the need to infer a hidden syntactic structure for sentences presented as word-sequences, based on the combination of bottom-up distributional learning with indirect and ambiguous semantic feedback.

Table 1 Example input and feedback representation for the original BabySRL system. For each training sentence (a), gold standard semantic feedback (b) provided true abstract role labels for each argument, and gold standard part-of-speech tagging provided true identification of the nouns and verbs in the sentence (c). Each noun was treated as an argument by the semantic-role classifier; in the input to this classifier, nouns were represented (features (d)) by the target argument and predicate themselves, and features indicating the position of each noun in a linear sequence of nouns (NPatttern or NPat, e.g., 1st of 2 nouns, 2nd of 2 nouns) and its position relative to the verb (VPosition or VPos). Section 4.1.1 will further describe these features

(a) Sentence		The	girl	tickled	the	boy
(b) Semantic feedback			A0			A1
(c) Syntactic structure			N	V		N
(d) Feature representation	*girl*	argument:girl predicate:tickled NPat: 1st of 2 Ns VPos:before verb		*boy*		argument:boy predicate:tickled NPat: 2nd of 2 Ns VPos: after verb

Much previous computational work has grappled with core questions about the earliest steps of language acquisition, and about links between verb syntax and meaning. Here we briefly review some major themes in these literatures, focusing on the range of assumptions made by various classes of models. In particular, we specify what problems of language learning each class of models attempts to solve, and what input and feedback assumptions they rely on to do so. As we shall see, the field has largely kept separate the learning of syntactic categories and structures on the one hand, and the learning of syntax-semantics links on the other. Few models attempt to combine the solutions to both of these problems, and we would argue that none simultaneously reflect the two central ambiguity problems (of sentence and scene input) that face the learner.

First, a large and varied class of computational models explores the use of distributional learning in a constrained architecture to permit the unsupervised identification of syntactic categories or structures. For example, clustering words based on similar distributional contexts (e.g., preceding and/or following words) results in word-classes that strongly resemble syntactic categories (e.g., [11, 30, 48, 62, 63]). In these systems, the text itself is typically the only input to the learner, but the nature of the classes also depends on the model's assumptions about how much context is available, and how (and how many) clusters are formed. Several influential recent models have extended such distributional analysis techniques to discover the constituent structure of sentences, and hierarchical dependencies between words or constituents (e.g. [7, 53, 81, 85]). These models again are unsupervised in the sense that they receive only word-sequences (or word-class sequences) as input, with no direct feedback about the accuracy of the structures they infer. They are also constrained by various assumptions about the nature of the structures to be uncovered (e.g., binary hierarchical structures), and by pressures toward generalization (e.g., minimum description length assumptions). The constraints imposed constitute

the model's fragment of Universal Grammar. These models inherit a long-standing focus on the importance of distributional analysis in linguistics [45,86]; jointly, such models demonstrate that appropriately constrained distributional analysis yields powerful cues to grammatical categories and structures. However, these models create *unlabeled* categories and structures, yielding no clear way to link their outputs into a grammar, or a system for interpreting or producing sentences. For the most part, distributional learning models have not been linked with models of sentence processing (though we will discuss one exception to this rule below). This is one of the goals of the current work, to link bottom-up distributional learning with a system for semantic-role labeling.

Second, a distinct class of models tackles the learning of relationships between syntax and semantics. A prominent recent approach is to use hierarchical Bayesian models to learn flexible, multi-level links between syntax and semantics, including syntactic-semantic classes of verbs and abstract verb constructions (e.g., [68, 69]), the abstract semantic roles that are linked with particular argument positions within verb frames or constructions [1], and verbs' selection restrictions [2]. These models address fascinating questions about the nature and representation of links between form and meaning. However, they do not attempt to address the ambiguity of either the sentence or scene input for the novice learner. Models in this class typically begin with input sentences that are already specified in both syntactic and semantic terms.

For example, Table 2 presents an input representation for the sentence "Sarah ate lunch" as presented to the models of [1, 2]. The syntactic part of this representation includes the identity of the verb, and the identity, number, and order of the verb's arguments. The semantic part is constructed based on hand-annotated verb usages and semantic properties extracted from the WordNet hierarchy [61]. The semantic representations provide both lexical-semantic features of the arguments and verb (e.g., that 'Sarah' is female) and features representing the role each argument plays in the event denoted by the verb (e.g., Sarah's role is volitional); the role features are derived from theoretical descriptions of the semantic primitives underlying abstract thematic roles (e.g., [28]). Thus, like the original BabySRL described above, models in this class represent a learner that has already acquired the grammatical categories and meanings of the words in the sentence, and can identify the relational meaning of the sentence. In essence, these models assume that identifying basic aspects of the syntactic structure of the sentence, and identifying the sentence's meaning, are separate problems that can be addressed as precursors to discovering links between syntax and semantics. The key argument of both the syntactic and semantic bootstrapping theories is that this is not true; on the contrary, links between syntax and semantics play a crucial role in allowing the learner to identify the syntax of the sentence, its meaning, or both [54, 70].

An influential model by Chang and colleagues is an exception to the rule that distributional-learning models are kept separate from higher-level language processing tasks. Chang et al. [14] implemented a model that learns to link syntax and semantics without predefined syntactic representations. Chang et al. modeled learning in a system that yokes a syntactic sequencing system consisting of a simple

Table 2 Example input sentence and extracted verb frame from [1]. The model learns to identify the subset of lexical and role features that are characteristic of each argument position within similar verb usages, thus learning abstractions such as 'agent' and 'patient'. The model assumes knowledge of the meanings of individual verbs and their arguments (using the WordNet hierarchy and hand-constructed event-role representations), and also syntactic knowledge of the identity of the verb and arguments, and the number and order of arguments in the sentence

(a) Sentence		Sarah ate lunch
(b) Syntactic pattern		arg1 verb arg2
(c) Semantic properties	verb:	{act, consume}
	eat	
	arg1	
	lexical:	{woman, adult female, female, person ...}
	role:	{volitional, affecting, animate ...}
	arg2	
	lexical:	{meal, repast, nourishment ...}
	role:	{non-independently exist, affected ...}

recurrent network (SRN), to a distinct message system that represents the meaning of each input sentence. The message system represents each sentence's meaning via lexical-semantic representations that specify what particular actions and entities are involved in the meaning, bound to abstract event-role slots (action, agent, theme, goal ...) that specify how many and what argument-roles are involved. In a typical training trial, the model is presented with a fixed message for the sentence, and a sequence of words conveying that message. The model tries to predict each next word in the sentence from the previous words, based on prior learning in the SRN and knowledge of the message. A key feature of this model is that the hidden units of the SRN are linked by learnable weights to the abstract event-role slots of the message system, but not to the lexical-semantic part of the message. This "Dual-Path" architecture keeps lexical-semantic information out of the syntactic sequencing system, thus ensuring that the model formulates abstract rather than word-specific syntactic representations in its hidden units. This system is unique in that it models online sentence processing, making predictions that change word by word as the sentence unfolds; thus, unlike the other models discussed in this section (including the BabySRL), it can be used to investigate how syntactic learning depends on the order in which information becomes available in sentences. The current effort shares with the dual-path model the linking of distributional learning into a system that learns to link syntax and semantics. However, the dual-path model creates syntactic representations by assuming the child already has accurate semantic representations of the input sentences. This model therefore resembles semantic bootstrapping in its reliance on meaning to drive syntax learning in a constrained architecture. We sought to create a model in which the problems of sentence and scene ambiguity could be solved jointly, allowing very partial syntactic constraints to help select a meaning from an ambiguous scene.

In jointly addressing these two types of ambiguity, our work could be viewed as analogous to a recent model of the task of word segmentation, which logically precedes the sentence-interpretation task we examine here. Johnson et al. [49] present a computational model that jointly learns word segmentation along with word-referent mappings; they demonstrate synergistic benefits from learning to solve these problems jointly. Here we try to apply a similar insight at a different level of analysis, to learn about the structure of the sentence (identifying arguments and predicates) along with a semantic analysis of the sentence (identifying semantic roles). These high-level processing steps of course also depend on word-segmentation success; although we do not yet incorporate this step, it could be argued that additional benefits could be achieved by learning jointly across all levels of language processing, from word segmentation through sentence-structure identification to semantic interpretation.

In learning semantic role labeling, it is well known that the parsing step which gives structure to the sentence is pivotal to final role labeling performance [39, 72]. Given the dependence of semantic role labeling on parsing accuracy, there is considerable interest in trying to learn syntax and semantics jointly, with two recent CoNLL shared tasks devoted to this problem [44,82]. In both cases, the best systems learned syntax and semantics separately, then applied them together, so at this level of language learning the promise of joint synergies has yet to be realized.

3 Model of Language Acquisition

As noted earlier, Semantic Role Labeling is an NLP task involving identifying and classifying the verbal predicate-argument structures in a sentence, assigning semantic roles to arguments of verbs. Combined with the development of robust syntactic parsers, this level of semantic analysis should aid other tasks requiring intelligent handling of natural language sentences, including information extraction and language understanding. A large literature exploring the SRL task began to emerge with the development of the PropBank semantic annotated corpora [52, 67] and the introduction of the CoNLL (Annual Conference on Computational Natural Language Learning) shared task SRL competitions [12, 13]. For a good review of the SRL task along with a summary of the state of the art, see [59].

To illustrate, (1) is a sentence from PropBank:

(1) Mr. Monsky *sees* much bigger changes ahead.

The SRL task is to identify the arguments of the verb "sees" and classify their roles in this structure, producing the labeling in (2). In example (2), square brackets mark the identified arguments; A0 (sometimes written as Arg-0) represents the agent, in this case the seer, A1 (also Arg-1) represents the patient, or that which is seen, and AM-LOC is an adjunct that specifies the location of the thing being seen.

(2) [$_{A0}$ Mr. Monsky] *sees* [$_{A1}$ much bigger changes] [$_{AM-LOC}$ ahead] .

PropBank defines two types of argument roles: core roles A0 through A5, and adjunct-like roles such as the AM-LOC above.[1] The core roles in the PropBank coding scheme represent a strong assumption about the nature of semantic roles [67]; this assumption is also a key assumption of the structure-mapping account. That is, the core role labels (especially A0 and A1) are assumed to be abstract semantic roles that are shared across verbs, although the precise event-dependent meanings of the roles depends on the verb. For example, the argument of each verb whose role is closest to a prototypical agent [28] is marked as A0; this would include the seer for 'see', the giver for 'give', and so forth. The argument whose role is closest to a prototypical patient is designated A1; this includes the thing seen for 'see', the thing given for 'give', and so forth. These role assignments are given for each verb sense in the frame files of PropBank. Each frame file has a different frame set for each sense of a verb that specifies and defines both the possible roles and the allowable syntactic frames for this verb sense. The across-verb similarity of roles sharing the same role-label is less obvious for the higher-numbered roles. For example, A2 is a source for 'accept', and an instrument for 'kick'.

3.1 CHILDES Training Data

One goal of the BabySRL project was to assess the usefulness of a proposed set of initial syntactic representations given natural corpora of child directed speech. Therefore we used as input samples of parental speech to three children (Adam, Eve, and Sarah; [10]), available via CHILDES [56]. The semantic-role-annotated corpus used in this project consists of parental utterances from sections Adam 01–23 (child age 2;3–3;2), Eve 01–20 (1;6–2;3), and Sarah 01–90 (2;3–4;1). All verb-containing utterances without symbols indicating disfluencies were automatically parsed with the Charniak parser [16] and annotated using an existing SRL system [72]; errors were then hand-corrected. The final annotated sample contains 15,148 sentences, 16,730 propositions, with 32,205 arguments: 3,951 propositions and 8,107 arguments in the Adam corpus, 4,209 propositions and 8,499 arguments in Eve, and 8,570 propositions and 15,599 arguments in Sarah.

3.1.1 Preprocessing and Annotation

During preprocessing of the CDS transcripts, only utterances from the Mother and Father were used. Other adults were typically present, including the researchers who collected the data, but we focused on parental speech because we considered it most

[1] In our corpus the full set of role labels is: A0, A1, A2, A3, A4, AM-ADV, AM-CAU, AM-DIR, AM-DIS, AM-EXT, AM-LOC, AM-MNR, AM-MOD, AM-NEG, AM-PNC, AM-PRD, AM-PRP, AM-RCL, AM-TMP.

likely to be typical CDS. Because our goal was to create a corpus for studying input for language learning, we made no attempt to annotate the children's speech.

In the process of annotation, as noted above we removed all parental utterances that contained symbols indicating unintelligible speech, or that did not contain a verb. In addition, after pilot annotation of utterances to one child (Eve), additional guidelines were set, especially in regard to what constituted a main or auxiliary verb. In particular, we decided not to annotate the verb 'to be' even when it was the main verb in the sentence. As a result of these decisions, although there were 45,166 parental utterances in the sections annotated, only 15,148 were parsed and annotated, fewer than 34 % of all utterances. This may seem like a surprisingly small proportion of the input to the children, but many of the ignored utterances were single-word exclamations ("Yes", "What?", "Alright," etc.), or were phrasal fragments that did not contain a main verb ("No graham crackers today." "Macaroni for supper?"). Such fragments are common in casual speech, and particularly so in speech to children. For example, in another corpus of child-directed English, only 52 % of the utterances were full clauses (the rest were phrasal fragments or single-word exclamations), and a substantial proportion of the full clauses had 'to be' as their main verb, as in "Who's so tall?" [33].

Annotators were instructed to follow the PropBank guidelines [67] in their semantic annotations, basing decisions on PropBank's previously-identified verb frames. If no frame existed for a specific verb (such as "tickle", found in CDS but not in the newswire text on which PropBank was developed), or a frame had to be modified to accommodate uses specific to casual speech, then the annotators were free to make a new decision and note this addition.[2]

In the main experiments reported in this chapter we used samples of parental speech to one child (Adam; [10]) as training and test data, sections 01–20 (child age 2;3–3;1) for training, and sections 21–23 for test. To simplify evaluation, we restricted training and testing to the subset of sentences with a single predicate (over 85 % of the annotated sentences). Additionally, in argument identification we focus on noun arguments, as will be described below. This omits some arguments that are not nouns (e.g., 'blue' in "Paint it blue."), and some semantic roles that are not typically carried by nouns. The final annotated sample contained 2,882 sentences, with 4,778 noun arguments.

3.2 Learning Model

The original architecture for our BabySRL was based on the standard pipeline architecture of a full SRL system [72], illustrated in the top row of Fig. 1. The stages

[2]Corpus, decision files and additional annotation information available at http.://cogcomp.cs.illinois.edu/~connor2/babySRL/

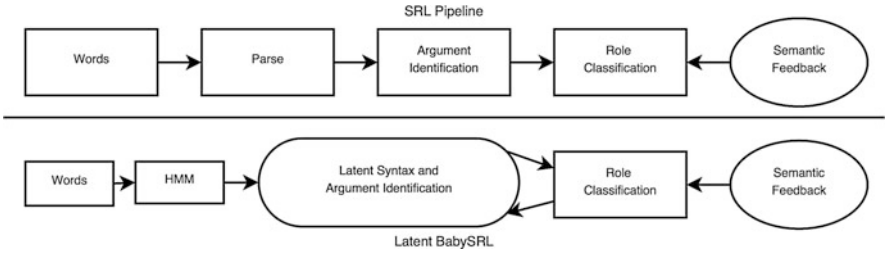

Fig. 1 Comparison of basic architecture of traditional pipeline approach for semantic role labeling versus latent BabySRL approach introduced here

are: (1) Parsing of the sentence, (2) Identifying potential arguments and predicates based on the parse, (3) Classifying role-labels for each potential argument, trained using role-labeled text. Each stage depended on the accuracy of the previous stages: argument identification depends on a correct parse, role labeling depends on correct arguments.

The key intuition of the Latent BabySRL described here is that we can use the task of semantic role labeling to generate and improve the intermediate syntactic representations that support this labeling. An SRL classifier determines the roles of arguments relative to a predicate in the sentence. The identity of arguments and predicates, of course, is not apparent in the surface form of the sentence. Therefore we suppose that this identification is part of a hidden structure for the sentence. The validity of this hidden structure determines the success of the semantic role labeling. As one of our assumptions about the starting-point of multi-word sentence comprehension, the semantic-role classifier assumes that nouns fill argument slots relative to verbs. Therefore the hidden structure that our system attempts to identify is a simple syntactic structure defined by identifying the nouns and verbs in the sentence, along with their linear order.

As shown in the bottom half of Fig. 1, the Latent BabySRL architecture roughly follows the standard pipeline, except that instead of a previously-trained syntactic parser and supervised argument identifier, we rely on an unsupervised clustering of words provided by a Hidden Markov Model (HMM), and a latent argument and predicate identifier that learns in response to feedback from the role classifier. In this system, decisions in the syntactic and semantic layers are linked together, and both are driven by semantic feedback from the world, given appropriate bottom-up information. The latent predicate and argument classifier learns what assists it in predicting semantic roles.

A similar HMM and the experiments in the next section were first presented in [23], and a preliminary version of the Latent BabySRL architecture first appeared in [24].

3.2.1 Unsupervised Part of Speech Clustering

As a first step in learning we used an unsupervised Hidden Markov Model (HMM) tagger to provide a context-sensitive clustering of words. We fed the learner large amounts of unlabeled text and allowed it to learn a structure over these data to ground future processing. This stage represents the assumption that the child is naturally exposed to large amounts of language, and will begin to gather distributional statistics over the input, independent of understanding the meaning of any words or sentences. Because we used transcribed speech, this step assumes that the learner can already correctly segment speech into words. The broader sample of text used to support this initial unsupervised HMM clustering came from child directed speech available in the CHILDES repository.[3] We again used only parents' sentences, and we removed sentences with fewer than three words or containing markers of disfluency. In the end we used 320,000 sentences from this set, including over two million word tokens and 17,000 unique words. Note that this larger HMM training set included the semantically tagged training data, treated for this purpose as unlabeled text.

The goal of this clustering was to provide a representation that allowed the learner to generalize over word forms. We chose an HMM because an HMM models the input word sequences as resulting from a partially predictable sequence of hidden states. As noted in Sect. 2, distributional statistics over word-strings yield considerable information about grammatical category membership; the HMM states therefore yield a useful unsupervised POS clustering of the input words, based on sequential distributional information, but without names for states. An HMM trained with expectation maximization (EM) is analogous to a simple process of predicting the next word in a stream and correcting connections accordingly for each sentence. We will refer to this HMM system as the HMM 'parser', even though of course parsing involves much more than part-of-speech clustering, largely because in the current version of the Latent BabySRL, the HMM-based clustering fills (part of) the role of the parser in the traditional SRL pipeline shown in Fig. 1.

An HMM can also easily incorporate additional knowledge during parameter estimation. The first (and simplest) HMM-based 'parser' we used was an HMM trained using EM with 80 hidden states. The number of hidden states was made relatively large to increase the likelihood of clusters corresponding to a single part of speech, while preserving some degree of generalization. Other researchers [47] have also found 80 states to be an effective point for creating a representation that is useful for further classification tasks, trading off complexity of training with specificity.

Johnson [48] observed that EM tends to create word clusters of uniform size, which does not reflect the way words cluster into parts of speech in natural languages. The addition of priors biasing the system toward a skewed allocation of words to classes can help. The second parser we used was an 80-state HMM trained

[3] We used parts of the Bloom [5,6], Brent [8], Brown [10], Clark [18], Cornell, MacWhinney [56], Post [26] and Providence [27] collections.

with Variational Bayes EM (VB) incorporating Dirichlet priors [3].[4] These priors assume one simple kind of innate knowledge on the learner's part, representing the expectation that the language will have a skewed distribution of word classes, with a relatively small number of large classes, and a larger number of small classes.

In the third and fourth parsers we experimented with enriching the HMM with other psycholinguistically plausible knowledge. Words of different grammatical categories differ in their phonological as well as in their distributional properties (e.g., [51, 64, 77]); thus combining phonological and distributional information improves the clustering of words into grammatical categories. The phonological difference between content and function words is particularly striking [77]. Even newborns can categorically distinguish content versus function words, based on the phonological difference between the two classes [78], and toddlers can use both phonology and frequency to identify novel words as likely content versus function words [46]. Human learners may treat content and function words as distinct classes from the start.

To implement this division into function and content words,[5] we started with a list of function word POS tags[6] and then found words that appeared predominantly with these POS tags, using tagged WSJ data [57]. We allocated a fixed number of states for these function words, and left the rest of the states for the content words. This amounts to initializing the emission matrix for the HMM with a block structure; words from one class cannot be emitted by states allocated to other classes. In previous work [23] we selected the exact allocation of states through tuning the heuristic system for argument and predicate identification examined in that work on a held-out set of CDS, settling on 5 states for punctuation, 30 states for function words, and 45 content word states. A similar block structure has been used before in speech recognition work [73], and this tactic requires far fewer resources than the full tagging dictionary that is often used to intelligently initialize an unsupervised POS classifier (e.g. [9, 74, 84]).

Because the function versus content word preclustering preceded HMM parameter estimation, it can be combined with either EM or VB learning. Although the initial preclustering independently forces sparsity on the emission matrix and allows more uniform sized clusters within each subset of HMM states, Dirichlet priors may still help, if word clusters within the function or content word subsets vary in size and frequency. Thus the third parser was an 80-state HMM trained with EM estimation, with 30 states pre-allocated to function words; the fourth parser was the same except that it was trained with VB EM.

[4] We tuned the priors using the same set of 8 value pairs suggested by Gao and Johnson [36], using a held out set of POS-tagged CDS to evaluate final performance. Our final values are an emission prior of 0.1 and a transitions prior of 0.0001; as a Dirichlet prior approaches 0 the resulting multinomial becomes peakier with most of the probability mass concentrated in a few points.

[5] We also include a small third class for punctuation, which is discarded.

[6] TO,IN,EX,POS,WDT,PDT,WRB,MD,CC,DT,RP,UH.

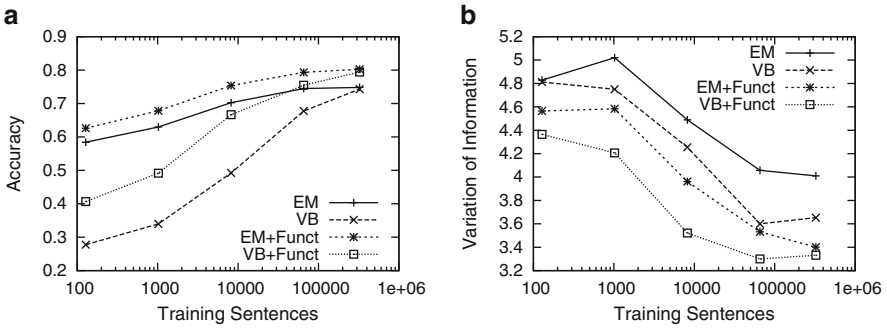

Fig. 2 Unsupervised Part of Speech results, matching states to gold-standard POS labels. All systems use 80 states, and are evaluated on a POS-labeled subset of CDS text, which comprises a subset of the HMM training data. Many-to-1 matching accuracy greedily matches states to their most frequent part of speech (Fig. 2a, higher is better). Variation of Information (Fig. 2b) is an information-theoretic measure summing mutual information between tags and states, proposed by Meilă [60], and first used for unsupervised part of speech in [41]. Smaller numbers are better, indicating less information lost in moving from the HMM states to the gold POS tags. Note that incorporating function word preclustering allowed both EM and VB algorithms to achieve the same performance with an order of magnitude fewer sentences. (**a**) Many to 1 accuracy. (**b**) Variation of information (Figure reproduced from [23])

3.2.2 HMM Evaluation

In previous work [23] we evaluated versions of these parsers (the first stage of our SRL system) on unsupervised POS clustering accuracy. Figure 2 shows the performance of the four parsers described above, using both many-to-one accuracy and variation of information to measure the match between fine-grained POS and the unsupervised parsers' decisions while varying the amount of text they were trained on. Each point on the graph represents the average result over ten training runs of the HMM with different samples of the unlabeled CDS.[7]

Many-to-one accuracy is an evaluation metric that permits multiple HMM states to map onto each POS tag: accuracy is measured by greedily mapping each state to the POS tag it most frequently occurs within the test data; all other occurrences of that state are then considered incorrect. EM can yield a better many-to-one score than VB-trained HMM [48], and our work showed the same result: across variations

[7]Note that the data shown in Fig. 2 reflect HMM initialization and training that differed slightly from that described in Sect. 3.2.1 and used in the experiments reported here: In that previous work, the set of function words differed slightly (e.g., in the current version we added 'not' to the function word set, and removed 'like' and 'have'), fewer states were allocated to punctuation (3 rather than 5), and the HMM was trained on a smaller sample of unlabeled text (up to 160,000 sentences rather than 320,000). The revised HMM parser used in the present experiments produced very similar results.

in amount of training data, EM yielded higher accuracy by this metric than VB, although these distinctions diminished as the amount of training data increased.

Variation of information is a metric of the distance between two clustering solutions (true POS labels and HMM states), which measures the loss and gain of information when moving from one clustering to the other. It is defined as $VI(C_1, C_2) = H(C_1|C_2) + H(C_2|C_1) = H(C_1) + H(C_2) - 2 * I(C_1, C_2)$, where $H(C)$ is the entropy of the clustering assignment C and $I(C_1, C_2)$ is the mutual information between the clustering C_1 and C_2. VI is a valid metric, and thus if two clusterings are identical, their VI will be 0.

These data show that the HMM yielded robust POS clustering, and that the four versions differed from each other in interesting ways. In particular, the content vs. function-word split improved POS clustering performance. Measured both by many-to-1 accuracy and VI, adding the function word split improved performance, for both EM and VB training. Thus a preclustering of content and function words, which we have argued is plausible for learners based on the well-established phonological differences between these classes, improves the automatic identification of POS clusters from text. In future sections we use the VB+Funct HMM, the best-performing system in this evaluation, as the first step in the Latent BabySRL. The HMM states both yield additional representational features that permit generalization across words, and give us a means of incorporating some minimally-supervised syntactic constraints on sentence interpretation.

4 Latent Training

Once a potential predicate and arguments have been identified (via latent training as described in this section), a role classifier must assign a semantic role to each argument relative to the predicate. The role classifier can only rely on features that can be computed with information available from previous stages of input processing, and from prior learning. The latent argument and predicate identifier is trained to best support accurate role classification. We trained this model in an online fashion in which we present each sentence along with some semantic constraints as feedback; both the semantic-role and the latent argument and predicate classifier then update themselves accordingly. In this section we will describe how the model is trained and what representations are used.

We can phrase our problem of Semantic Role Labeling as learning a structured prediction task, which depends on some latent structure (argument and predicate identification). As input we have the sequence of words and HMM states for a given sentence, and the output is a role-labeled predicate-argument structure. The goal in our structured prediction task is to learn a linear function $f_w : X \rightarrow Y$ that maps from the input space X (sentences) to output space Y (role labeled argument structure):

$$f_w(x) = \arg\max_{y \in Y} \max_{h \in H} w \cdot \Phi(x, h, y) \tag{1}$$

Here H is a space of hidden latent structures that describes some connection between X and Y (identification of arguments and predicate), Φ is a feature encoding for the complete role labeled X, H, Y example structure, w is the learned weight vector that scores structures based on their feature encoding, and both $w, \Phi \in \mathbb{R}^n$.

Conventionally the weight vector w would be learned from a set of labeled training examples $(x_i, y_i) \in X \times Y$, attempting to maximize the difference between the score for true structures y_i and all other structures for every training example. As we argued in the introduction to this chapter, it is implausible for the learner to receive veridical sentence meanings for each sentence (the set of role labels linked with arguments) as feedback for learning. The referential contexts that accompany speech are assumed to be ambiguous; this is the problem that syntactic bootstrapping sets out to solve. Therefore, instead of assuming that the learner is provided with a single true interpretation, we rephrase the learning problem such that for each sentence the learner is provided with a set of possible interpretations $Y_i \subseteq Y$ along with constraints on possible hidden structures $H_i \subseteq H$. In the next section we will describe specific implementations of this feedback scheme. However, in this section, for clarity in describing our algorithm and feature-sets, we use as an example the simplest case, in which only the true interpretation is provided.

Because of the max over H in the definition of f_w, the general optimization problem for finding the best w (in terms of minimizing a loss, or maximizing the margin between the true structure and all others given a training set of $\{x_i, y_i\}_{i=1}^{M}$ labeled examples) is non-convex. Previously this has been solved using some variant of latent structure optimization [15, 87]. Here we used an online approach and a modification of Collin's Structured Perceptron [20] with margin [50]. This basic, purely latent algorithm (Algorithm 2) uses an approximation employed in [17, 31] where for each example the best h^* is found (according to the current model and true output structure) and then the classifier is updated using that fixed structure. In this algorithm C is a fixed margin (set at 1.0) that must seperate the true structure from the next highest prediction for the algorithm to not modify the weight vector ($\mathbf{1}[y \neq y_i^*]$ is an indicator function that is 1 for all y that are not the true structure). The constant α_w represents the learning rate.

The intuition behind Algorithm 2 is that for every sentence the learner knows the true meaning, or set of meanings that contain the true meaning (Y_i), so it is able to find the arrangement of arguments and predicate (hidden structure h^*) that best supports that meaning according to what it has already learned (current weight vector w_t). Once we identify the latent arguments and predicate, we use this identification to update the weight vector so the true role prediction y_i^* will be more likely in the future (line 5 and 6, structured perceptron update).

As stated in Algorithm 2, h^*, the best set of arguments and predicates, is found and then forgotten for each input sentence x. If we are interested in h beyond its application to learning the weights w to predict semantic roles y, such as for generalizing to better find the arguments and predicate in related sentences x, then we need a method for storing this information and passing it on to new examples.

Algorithm 2 Purely latent structure perceptron

1: Initialize $w_0, t = 0$
2: **repeat**
3: **for all** Sentences (x_i, Y_i) **do**
4: $(h_i^*, y_i^*) \leftarrow \arg\max_{h \in H_i, y \in Y_i} w_t \cdot \Phi_w(x_i, h, y)$
5: $y' \leftarrow \arg\max_y w_t \cdot \Phi_w(x_i, h_i^*, y) + C * \mathbf{1}[y \neq y_i^*]$
6: $w_{t+1} \leftarrow w_t + \alpha_w(\Phi_w(x_i, h_i^*, y_i^*) - \Phi_w(x_i, h_i^*, y'))$
7: $t \leftarrow t + 1$
8: **end for**
9: **until** Convergence

Algorithm 3 Online latent classifier training

1: Initialize $w_0, u_0, t = 0$
2: **repeat**
3: **for all** Sentences (x_i, Y_i) **do**
4: $(h_i^*, y_i^*) \leftarrow \arg\max_{h \in H_i, y \in Y_i} w_t \cdot \Phi_w(x_i, h, y) + u_t \cdot \Phi_u(x_i, h)$
{Update u to predict h^*}
5: $h' \leftarrow \arg\max_h u_t \cdot \Phi_u(x_i, h) + C * \mathbf{1}[h \neq h_i^*]$
6: $u_{t+1} \leftarrow u_t + \alpha_u(\Phi_u(x_i, h_i^*) - \Phi_u(x_i, h'))$
{Update w based on h^* to predict y^*}
7: $y' \leftarrow \arg\max_y w_t \cdot \Phi_w(x_i, h_i^*, y) + C * \mathbf{1}[y \neq y_i^*]$
8: $w_{t+1} \leftarrow w_t + \alpha_w(\Phi_w(x_i, h_i^*, y_i^*) - \Phi_w(x_i, h_i^*, y'))$
9: $t \leftarrow t + 1$
10: **end for**
11: **until** Convergence

To solve this problem, we trained a latent predicate and argument classifier along with the role classifier, such that during the latent prediction for each sentence we find the structure that maximizes the score of both role classification and structure prediction. This algorithm is summarized in Algorithm 3. The end result is two classifiers, f_u to predict hidden structure and f_w to use the hidden structure, that have been trained to work together to minimize semantic-role classification training error.

The intuition behind Algorithm 3 is that for each sentence the learner finds the best joint meaning and structure based on the current classifiers and semantic constraints (line 4), then seperately updates the latent structure f_u and output structure f_w classifiers given this selection. In the case where we have perfect high level semantic feedback $Y_i = y_i$, the role classifier will search for the argument structure that is most useful in predicting the correct labels. More generally, partial feedback, which constrains the set of possible interpretations but does not indicate the one true meaning, may be provided and used for both labeling Y_i and hidden structure H_i.

This learning model allows us to experiment with the trade-offs among different possible sources of information for language acquisition. Given perfect or highly informative semantic feedback, our constrained learner can fairly directly infer the true argument(s) for each sentence, and use this as feedback to train the latent

argument and predicate identification (what we might term semantic bootstrapping). On the other hand, if the semantic role feedback is loosened considerably so as *not* to provide information about the true number or identity of arguments in the sentence, the system cannot learn in the same way. In this case, however, the system may still learn if further constraints on the hidden syntactic structure are provided through another route, via a straight-forward implementation of the structure-mapping mechanism for early syntactic bootstrapping.

4.1 Argument, Predicate and Role Classification

For the latent structure training method to work, and for the hidden structure classifier to learn, the semantic role classifier and feature set (f_w and Φ_w respectively) must make use of the hidden structure information h. In our case, the role classifier makes use of (and thus modifies during training) the hidden argument and predicate identification in two ways. The first of these is quite direct: semantic role predictions are made relative to specific arguments and predicates. Semantic-role feedback therefore provides information about the identity of the nouns in the sentence. The second way in which the role classifier makes use of the hidden argument and predicate structure is less direct: The representations used by the SRL classifier determine which aspects of the predictions of the argument and predicate latent classifier are particularly useful in semantic role labeling, and therefore change via the learning permitted by indirect semantic-role feedback.

In the simplest case we use the full set of correct role labels as feedback. We implement this by providing correct labels for each word in the input sentence that was selected by the latent classifier as an argument and is the head noun of an argument-phrase. Thus the optimal prediction by the argument classifier will come to include at least those words. The predicate classifier will therefore learn to identify predicates so as to maximize the accuracy of SRL predictions given these arguments. This represents the case of semantically-driven learning where veridical semantic feedback provides enough information to drive learning of both semantics and syntax. With more ambiguous semantic feedback, the hidden argument and predicate prediction is not directed by straightforward matching of a full set of noun arguments identified via semantic feedback. Nonetheless, the system is still driven to select a hidden structure that best allows the role classifier to predict with what little semantic constraint is provided. Without further constraints on the hidden structure itself, there may not be enough information to drive hidden structure learning.

In turn, the hidden structure prediction of arguments and predicate depends on the words and HMM states below it, both in terms of features for prediction and constraints on possible structures. The hidden argument and predicate structure we are interested in labels each word in the sentence as either an argument (noun), a predicate (verb), or neither. We used the function/content word state split in the HMM to limit prediction of arguments and predicates to only those words identified

as content words. In generating the range of possible hidden structures over content words, the latent structure classifier considers only those with exactly one predicate and one to four arguments.

As an example take the sentence "She likes yellow flowers." There are four content words; with the constraint that exactly one is a predicate and at least one is an argument, there are 28 possible predicate/argument structures, including the correct assignment where 'She' and 'flowers' are arguments of the predicate 'likes.' The full semantic feedback would indicate that 'She' is an agent and 'flowers' is a patient, so the latent score the SRL classifier predicts (line 4 in Algorithm 3) will be the sum of the score of assigning agent to 'She' and patient to 'flowers', assuming both those words are selected as arguments in h. If a word does not have a semantic role (such as non-argument-nouns 'likes' or 'yellow' here) then its predictions do not contribute to the score. Through this mechanism the full semantic feedback strongly constrains the latent argument structure to select the true argument nouns. Table 3 shows the two possible interpretations for "She likes yellow flowers." given full semantic feedback that identifies the roles of the correct arguments. Decisions regarding 'likes' and 'yellow' must then depend on the representation used by both the latent-structure predicate identifier and semantic-role classifier.

4.1.1 Features

For the semantic-role classifier we started with the same base BabySRL features developed in [21], simple structures that can be derived from a linear sequence of candidate nouns and verb. These features include 'noun pattern' features indicating the position of each proposed noun in the ordered set of nouns identified in the sentence (e.g., first of three, second of two, etc; NPat in Table 3), and 'verb position' features indicating the position of each proposed noun relative to the proposed verb (before or after; VPos in Table 3). In the above example, given the correct argument assignment, these features would specify that 'She' is the first of two nouns and 'flowers' is the second of two. No matter whether 'likes' or 'yellow' is selected as a predicate, 'She' is before the verb and 'flowers' is after it. In addition, we used a more complicated feature set that includes NPat and VPos features along with commonly-used features such as the words surrounding each proposed noun argument, and conjunctions of NPat and VPos features with the identified predicate (e.g., the proposed predicate is 'likes' and the target noun is before the verb); such features should make the role classifier more dependent on correct predicate identification.

For the argument and predicate structure classifiers the representation $\Phi_u(x, h)$ only depends on words and the other arguments and predicates in the proposed structure. Each word is represented by its word form, the most likely HMM state given the entire sentence, and the word before and after. We also specified additional features specific to argument or predicate classification: the argument classifier uses noun pattern (NPat in Table 3), and the predicate representation uses the conjunction

Table 3 Example Sentence, showing (a) the full (gold standard) semantic feedback that provides true roles for each argument, but no indication of the predicate, as well as (b) two possible hidden structures given this level of feedback. The next rows show (c) the feature representations for individual words. The Semantic Feature set shows the feature representation of each argument as used in SRL classification; the Structure Feature set shows the feature representation of the first argument and the predicate in two of the 28 possible hidden structures. See text Sect. 4.1.1 for further description of the features

(a) Sentence				She likes yellow flowers		
full feedback				A0		A1
(b) Possible interpretation 1				Possible interpretation 2		
Sentence	she	likes	yellow flowers	Sentence	she	likes yellow flowers
argument struct.	N	V	N	argument struct.	N	V N
(c) Feature representation				Feature representation		
Semantic feat. $\Phi_w(x, h, y)$	she		argument:she predicate:likes NPat: 1 of 2 VPos:Before Verb w+1:likes	Semantic feat. $\Phi_w(x, h, y)$	she	argument:she predicate:yellow NPat: 1 of 2 VPos:Before Verb w+1:likes
	flowers		argument:flowers predicate:likes NPat: 2 of 2 VPos: After Verb w−1:yellow w+1:.		flowers	argument:flowers predicate:yellow NPat: 2 of 2 VPos: After Verb w−1:yellow w+1:.
Structure feat. $\Phi_u(x, h)$	she=N		word:she hmm:35 verb:likes w+1:likes hmm+1:42 NPat: 1 of 2	Structure feat. $\Phi_u(x, h)$	she=N	word:she hmm:35 verb:yellow w+1:likes hmm+1:42 NPat: 1 of 2
	likes=V		verb:likes hmm:42 w−1:she hmm−1:35 w+1:yellow hmm+1:57 v:likes&2 args suffixes: s,es,kes		yellow=V	verb:yellow hmm:57 w−1:likes hmm−1:42 w+1:flowers hmm+1:37 v:flowers&2 args suffixes: w,ow,low

of the verb and number of arguments (e.g., 'v:likes & 2args' in Table 3), as well as all suffixes of length up to three as a simple verb ending feature.[8]

[8]This roughly represents phonological/distribution information that might be useful for clustering verbs together (e.g., [64]), but that is not exploited by our HMM because the HMM takes transcribed words as input.

It should be noted that both the purely latent (Algorithm 2) and latent classifier we have been discussing (Algorithm 3) require finding the max over hidden structures and labelings according to some set of constraints. As implemented with the sentences found in our child directed speech sample, it is possible to search over all possible argument and predicate structures. In our set of training sentences there were at most nine content words in any one sentence, which requires searching over 1,458 structures of exactly one predicate and at most four arguments. On average there were only 3.5 content words a sentence. Once we move to more complicated language an alternative approximate search strategy will need to be employed.

5 Experimental Evaluation

To evaluate the Latent BabySRL, we examined both how well the final role classifier performed, and how accurately the latent predicate and argument classifiers identified the correct structures when trained with only indirect semantic feedback. Because in our training sentences there was only one true predicate per sentence, we report the predicate accuracy as the percentage of sentences with the correct predicate identified. For the identification of noun arguments, because there were multiple possible predictions per sentence, we report F1: the harmonic mean of precision and recall in identifying true arguments. Likewise, in evaluating semantic-role classification, because there were many possible role labels and arguments to be labeled, we report the overall semantic role F1 over all arguments and label predictions.[9]

Our first experiment tested online latent training with *full semantic feedback*. To provide an upper bound comparison we trained with perfect argument knowledge, so in this case both classifiers were fully and separately supervised (Gold Arguments in Table 4). This upper bound reflects the levels of argument-identification and SRL performance that are possible given our simple feature set and child-directed sentence corpus. As a lower bound comparison for predicate-argument classification we also include the expected result of selecting a random predicate/argument structure for each sentence (Random Arguments in Table 4).

Table 4 shows the performance of the two algorithms from Sect. 4 compared to the just-mentioned upper and lower bounds. All classifiers used the full feature sets from Sect. 4.1. Recall that the purely latent method (Algorithm 2) did not use an intermediate latent structure classifier, so it selected arguments and predicates only to maximize the role classifier prediction for the current sentence. In contrast, incorporating a latent classifier into the training (Algorithm 3) yielded a large boost in both argument and predicate identification performance and final role performance. Thus, given full semantic feedback, the argument and predicate classifier

[9]Because we focus on noun arguments, we miss those predicate arguments that do not include any nouns; the maximum SRL role F1 with only noun arguments correct is 0.8255.

Table 4 Results on held-out test set of SRL with arguments/predicate as latent structure, provided with full semantic feedback. With gold arguments, both the structure classifier and the role classifier are trained with full knowledge of the correct arguments for each sentence. Purely Latent does not use the latent argument and predicate classifier; it selects a structure for each sentence that maximizes role classification of true labels during training (Algorithm 2). Latent classifier training trains an argument/predicate identifier using the structure that the role classifier considers most likely to give the correct labeling (where we know correct labels for each noun argument), Algorithm 3

Training	Predicate %	Argument F1	Role F1
Gold arguments	0.9740	0.9238	0.6920
Purely latent	0.5844	0.6992	0.5588
Latent classifier	0.9263	0.8619	0.6623
Random arguments	0.3126	0.4580	–

effectively generalized the training signal provided by the latent semantic feedback to achieve nearly the performance of being trained on the true arguments explicitly (Gold Arguments). Of special note is the predicate identification performance; while full semantic feedback implicitly indicates true arguments, it says nothing about the true predicates. The predicate classifier was able to extract this information solely based on identifying latent structures that helped the role classifier make the correct role predictions.

As mentioned in Sect. 4.1, our algorithm depends on two kinds of representations: those that feed semantic role classification, and those that feed the hidden argument and predicate classifier. To investigate the interaction between the two classifiers' (hidden structure and SRL) representation choices, we tested the latent classifier with the full argument and predicate feature sets when the role classifier incorporated four different feature sets of increasing complexity: only the words identified as candidate nouns and verb (Words in Table 5), words plus noun pattern features (+NPat), the previous plus verb position features (+VPos), and a full model containing all these features as well as surrounding words and predicate conjunctions. With the addition to the SRL classifier of features that depend on more accurate latent structure identification, we should see improvements in both final role accuracy and argument and predicate identification. This experiment again used full role feedback.

Table 5 shows increasing performance with the increasing feature complexity of the semantic role classifier. Most notable is the large difference in predicate identification performance between those feature sets that heavily depend on accurate predicate information (+VPos and the full feature set in Table 5) and those that only use the word form of the identified predicate as a feature. In contrast, argument identification performance varied much less across feature sets in this experiment, because full semantic feedback always implicitly drives accurate argument identification. The increase in role classification performance across feature sets can be attributed both to a useful increase in representations used for SRL classification, and to the increased argument and predicate structure accuracy during both SRL training and testing. The relatively high level of SRL performance

Table 5 With full role feedback and latent classifier training, the role classifier features interact with the latent predicate-argument structure classifier. Better role classification through improved feature representation feeds back to allow for improved argument and predicate identification. The last two feature sets make strong use of the identity of the predicate, which encourages the predicate classifier to accurately identify the predicate. Each result represents the average over ten runs with random training order; numbers in parenthesis are standard deviations

Role features	Predicate %	Argument F1	Role F1
Words	0.64 (0.02)	0.81 (0.00)	0.63 (0.01)
+NPat	0.73 (0.05)	0.81 (0.00)	0.62 (0.01)
+VPos	0.93 (0.04)	0.83 (0.03)	0.65 (0.01)
+Surrounding words and predicate conjunctions	0.93 (0.03)	0.86 (0.04)	0.66 (0.01)

given the lexical features alone in Table 5 reflects the repetitive character of the corpus from which our training and test sentences were drawn: Given full semantic feedback, considerable success in role assignment can be achieved based on the argument-role biases of the target nouns (e.g., 'she', 'flowers') and the familiar verbs in our corpus of child-directed speech.

The results in this section show that the latent argument and predicate classifier, equipped with simple representations of the proposed sentence structure, can recruit indirect semantic-role feedback to learn to improve its representation of sentence structure, at least when given fully accurate semantic-role feedback. This result makes sense: the identity and position of the verb are useful in identifying the semantic roles of the verb's arguments; therefore the latent predicate-argument classifier could use the indirect semantic feedback to determine which word in the sentence was the verb. The full semantic-role feedback provided true information about the number and identity of arguments in each sentence; in the next section we take the crucial next step, reducing the integrity of the semantic role feedback to better reflect real-world ambiguity.

5.1 Ambiguous Semantic Feedback

The full semantic feedback used in the previous experiments, while less informative than absolute gold knowledge of true arguments and predicates, is still an unreasonable amount of feedback to grant a child first trying to understand sentences. The semantic feedback in our model represents the child's inference of sentence meaning from observation of the referential context. Because an ordinary scene makes available a number of possible objects, relations and semantic roles that might be mentioned, the child must learn to interpret sentences without prior knowledge of the true argument labels for the sentence or of how many arguments are present.

We implement this level of feedback by modifying the constraining sets H_i and Y_i used in line 4 of Algorithm 3. By loosening these sets we still provide feedback to

restrict the search space (thus modeling language learning as a partially supervised task, informed by inferences about meaning from scene observation), but not a veridical role-labeling for each sentence.

We tested two levels of reduced role feedback. The first, which we call Set of Labels, provides as feedback the true role labels that are present in the sentence, but does not indicate which words correspond to each role. In this case Y_i is just the set of all labelings that use exactly the true labels present, and H_i is constrained to be only those syntactic predicate-argument structures with the correct number of arguments. This feedback scheme represents a setting where the child knows the semantic relation involved, but either does not know the nouns in the sentence, or alternatively does not know whether the speaker meant 'chase' or 'flee' (and therefore cannot fix role order). To illustrate, given the sentence "Sarah chased Bill", Set of Labels feedback would indicate only that the sentence's meaning contains an agent and a patient, but not which word in the sentence plays which role.

Even this Set of Labels feedback scheme specifies the number of true arguments in the sentence. We can go a step further, supplying for each sentence a superset of the true labels from which the learner must select a labeling. In the Superset feedback case, Y_i includes the true labels, plus random additional labels such that for every sentence there are four labels to choose from, no matter the number of true arguments. Given Superset feedback, the learner is no longer constrained by the true number of arguments provided via semantic feedback, so must search over all argument structures and role labelings that come from some subset of the feedback set Y_i. This represents a setting in which the learner must select a possible interpretation of the sentence from a superset of possible meanings provided by the world around them. In the "Sarah chased Bill" example, the feedback would be a set of possible labels including the true agent and patient roles, but also two other roles such as recipient or location, and thus no scene-derived indication of how many of these roles are part of the sentence's meaning. This may seem an extreme reduction of the validity of semantic-role feedback. However, consider the following example: a careful analysis of video transcripts of parents talking to toddlers found that the parents were about equally likely to use intransitive motion verbs (e.g., 'go in') as transitive ones (e.g., 'put in') when describing events in which an agent acted on an object [75]. Evidently the presence of an agent in an event does not demand that a speaker choose a verb that encodes the agent's role. Similarly, in our earlier 'yellow flower' example, under many circumstances the speaker presumably could have said 'yellow flowers are nice' rather than 'she likes yellow flowers.' These considerations, and the 'human simulation' experiments described in Sect. 1.1 [40], all suggest that the number and identity of arguments in the speaker's intended meaning is not readily inferrable from world events without some guidance from the sentence.

As seen in Table 6, Set and Superset feedback seriously degrade performance compared to full role feedback. With superset feedback the learner cannot get a good foothold to begin correctly identifying structure and interpreting sentences, so its argument and predicate identification accuracy is little better than random. This suggests that information about the number and identity of arguments might

Table 6 Results when the amount of semantic feedback is decreased. Each value represents the mean over 20 training runs with shuffled sentence order; the numbers in parenthesis are the standard deviations. Full label feedback provides true role feedback for each noun. Set of Labels feedback provides an unordered set of true labels as feedback, so the learner must pick a structure and label assignment from this set. Superset goes one step further and provides a superset of labels that includes the true labels, so the learner does not know how many or which roles are mentioned in the sentence. With these ambiguous feedback schemes the classifiers are barely able to begin interpreting correctly, and with superset feedback the argument and predicate accuracy is only slightly better than random

Feedback	Pred %	Arg F1	A0	A1	Role F1
Full labels	0.94 (0.02)	0.89 (0.02)	0.85 (0.02)	0.75 (0.02)	0.64 (0.02)
Set of labels	0.40 (0.23)	0.62 (0.14)	0.47 (0.28)	0.38 (0.17)	0.34 (0.14)
Superset	0.35 (0.20)	0.57 (0.11)	0.46 (0.27)	0.33 (0.13)	0.29 (0.11)
Random	0.31	0.46			

be a necessary constraint in learning to understand sentences. In principle this information could be derived either from observation of scenes (assuming the child has access to some non-linguistic source of evidence about whether the speaker meant 'chase' or 'flee', 'put in' or 'go in') or from observation of sentences; the latter source of information is the essence of syntactic bootstrapping, as we discuss next.

6 Recovering Argument Knowledge

Considerable psycholinguistic evidence, reviewed briefly in Sect. 1.1, suggests that children learn some nouns before they start to interpret multi-word sentences, and thus some noun knowledge is available to scaffold the beginnings of sentence interpretation (e.g., [40]). This is syntactic bootstrapping, using structural features of the sentence to guide interpretation under ambiguity. If we can combine this extra source of knowledge with the Superset feedback described above, then perhaps the result will be enough information for the system to learn to identify nouns and verbs in sentences, and to classify the roles those nouns play.

Taking inspiration from the 'structure-mapping' account of syntactic bootstrapping, we model this starting point by attempting to identify nouns in each input sentence in a bottom-up, minimally-supervised manner. Once we know the number and identity of nouns in the sentence, this additional constraint on the hidden structure may allow the learner to overcome the semantic ambiguity introduced by Superset feedback. In the next section we will describe how we identify potential arguments using the distributional clustering provided by the HMM and a small seed set of concrete nouns. A similar form of this bottom-up argument-identification procedure was described in [23].

The bottom-up, minimally-supervised argument identifier we describe here addresses two problems facing the human learner. The first involves clustering

words by part-of-speech. As described in Sect. 3.2.1, we use a fairly standard Hidden Markov Model (HMM), supplemented by an a priori split between content and function words, to generate clusters of words that occur in similar distributional contexts. The second problem is more contentious: Having identified clusters of distributionally-similar words, how do children figure out what role these clusters of words play in a sentence interpretation system? Some clusters contain nouns, which are candidate arguments; others contain verbs, which take arguments. How is the child to know which are which?

The latent training procedure described in this chapter, when given full semantic feedback, accomplishes argument and predicate identification roughly by semantic bootstrapping: To return to our 'She likes yellow flowers' example, if the learner knows based on semantic feedback that 'she' is an agent, then the latent classifier learns to treat 'she' as a noun argument; the provision of abstract HMM-based features and noun-pattern features to the argument identification classifier permits it to generalize this learning to other words in similar sentence positions. But this use of semantic-role feedback to identify nouns as such seems counter-intuitive. In this section we spell out a simpler way to use a small step of semantic bootstrapping to automatically label some of the distributionally-derived clusters produced by the HMM tagger as nouns, thereby improving argument-identification from the bottom up, without requiring accurate semantic-role feedback.

6.1 Bottom-Up Argument Identification

The unsupervised HMM parser provides a state label for each word in each sentence; the goal of the argument identification stage is to use these states to label words as potential arguments, predicates or neither. As described in Sect. 1.1, the structure-mapping account of early syntactic bootstrapping holds that sentence comprehension is grounded in the learning of an initial set of nouns. Children are assumed to identify the referents of some concrete nouns via cross-situational learning [40, 79]. Children then assume, given the referential meanings of these nouns, that they are candidate arguments. Again, this involves a small step of semantic bootstrapping, using the referential semantics of already-learned words to identify them as nouns. We used a small set of known nouns to transform unlabeled word clusters into candidate arguments for the SRL: HMM states that occur frequently with known names for animate or inanimate objects are assumed to be argument states.

Given text parsed by the HMM parser and a seed list of known nouns, the argument identifier proceeds as illustrated in Algorithm 4. Algorithm 4 identifies noun states simply by counting the number of times each state is seen with a known noun ($freq_N(s)$ in Algorithm 4) in some HMM tagged text (Adam training data). Any state that appears at least four times with words from the seed noun list is identified as a noun state. Whenever these states are encountered in the future, the word associated with them, even if unknown, will be interpreted as a potential

Algorithm 4 Argument state identification

```
INPUT: Parsed Text T = list of (word, state) pairs
       Set of concrete nouns N
OUTPUT: Set of argument states A
A ← ∅
{Count Appearance of each state with a known noun}
freq_N(s) ← |{(w, s) ∈ T|w ∈ N}|
for all Content States s do
    if freq_N(s) ≥ 4 then
        Add s to A
    end if
end for
return A
```

argument. This use of a seed list with distributional clustering is similar to Prototype Driven Learning [43], except in the present case we provide information on only one class. A similar approach was proposed by Mintz [62], using semantic knowledge of a small set of seed nouns to tag pre-existing distributionally-based clusters as noun clusters.

Because we train our HMM with a preclustering of states into function and content words, we use this information in the minimally supervised argument identification. Only content word states are considered to be potential argument states, thus eliminating any extraneous function words from consideration. This of course improves identification performance, because it only eliminates potential errors.

To generate a plausible 'seed' set of concrete nouns, we used lexical development norms [25], selecting all words for things or people that were commonly produced by 20-month-olds (over 50 % reported), and that appeared at least five times in our training data. Because this is a list of words that children produce, it represents a lower bound on the set of words that children at this age should comprehend. This yielded 71 words, including words for common animals ('pig', 'kitty', 'puppy'), objects ('truck', 'banana', 'telephone'), people ('mommy', 'daddy'), and some pronouns ('me' and 'mine'). To this set we added the pronouns 'you' and 'I', as well as given names 'adam', 'eve' and 'sarah'. The inclusion of pronouns in our list of known nouns represents the assumption that toddlers have already identified pronouns as referential terms. Even 19-month-olds assign appropriately different interpretations to novel verbs presented in simple transitive versus intransitive sentences with pronoun arguments ("He's kradding him!" vs. "He's kradding!"; [88]).

The resulting set of 76 seed nouns represents a high-precision set of argument nouns: They are not highly frequent in the data (except for the pronouns), but they nearly always appear as nouns and as arguments in the data (over 99 % of the occurrences of words in this list in our training data are nouns or pronouns, over 97 % are part of arguments). Given this high precision, we set a very permissive condition that identifies argument states as those HMM states that appear four or more times with known seed nouns. In our experiments we set the threshold

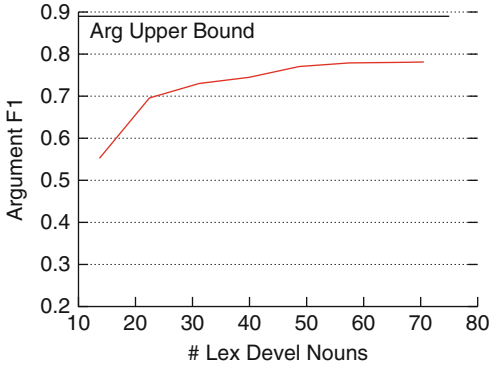

Fig. 3 Effect of number of concrete nouns for seeding argument identification. To generate these results, one HMM trained with VB+Funct was selected out of ten models with different random initializations (the best in terms of lowest perplexity on large corpus of untagged training data). Adam training data was then tagged with states from this HMM; for each lexical development seed set size, 100 runs with random selection of seed nouns were averaged together to produce the data shown here. Argument identification accuracy is computed for the Adam data set using true argument boundaries from hand labeled data. With 76 seed nouns, the argument identifier achieves nearly 0.80 F1

of known nouns appearing with an HMM state to four through tuning argument identification on a held-out set of argument-identified sentences.

6.1.1 Argument Identification Evaluation

Figure 3 shows the argument identification accuracy of the minimally supervised argument identification system, with increasing numbers of seed nouns sampled from the set of 76. First, using the HMM model with the function-content-word split and VB training, we generated 10 models over the large untagged HMM training corpus with different random initializations, and selected the one with the lowest perplexity (highest log likelihood) for use in the argument identification experiments. We then tagged the Adam SRL training data with states from this selected HMM, using the state for each word that had highest marginal probability given the rest of the sentence (using forward-backward algorithm). In this figure, for each set size of seed nouns, we report the mean over 100 runs of the argument identification with a random selection of seed nouns in each run.

We evaluated this performance compared to hand labeled data with true argument and predicate boundaries. We present the primary argument (A0–4) identification accuracy using the F1 measure, with precision calculated as the proportion of identified arguments that appear as part of a true argument, and recall as the proportion of true arguments that cover some state identified as an argument. This is a rather lenient measure of accuracy since we are comparing identified individual words to full phrase boundaries.

Table 7 Results when the amount of semantic feedback is decreased, but bottom-up syntactic information is used to help constrain the possible hidden structures and recover from ambiguous semantic feedback. The top three rows of data are reproduced from Table 6. We introduce extra information by constraining the possible argument structures for each training example using syntactic knowledge, either bottom-up from an HMM-based minimally supervised argument identifier, or via knowledge of true arguments. Once extra information about argument identity is introduced, whether true arguments or the HMM-identified arguments, the learner is able to make use of the Superset feedback, and begin to identify the agent and patient roles (A0 and A1), and the predicate

Feedback	Pred %	Arg F1	A0	A1	Role F1
Full labels	0.94(0.02)	0.89(0.02)	0.85(0.02)	0.75(0.02)	0.64(0.02)
Set of labels	0.40(0.23)	0.62(0.14)	0.47(0.28)	0.38(0.17)	0.34(0.14)
Superset	0.35(0.20)	0.57(0.11)	0.46(0.27)	0.33(0.13)	0.29(0.11)
Superset + HMM args	0.87(0.10)	0.88(0.01)	0.68(0.25)	0.54(0.16)	0.48(0.15)
Superset + true args	0.86(0.09)	0.92(0.01)	0.69(0.21)	0.61(0.13)	0.52(0.13)
Superset + true args & Pred	0.97(0.00)	0.93(0.00)	0.68(0.19)	0.61(0.12)	0.52(0.11)
Random	0.31	0.46			

As Fig. 3 shows, this minimally-supervised argument identification system can successfully identify arguments starting with a handful of concrete nouns. Even with just 10 nouns, argument identification is almost 0.6 F1; with 76 nouns (still a modest number relative to toddlers' estimated comprehension vocabularies), argument identification improves to nearly 0.8 F1. Even so, we have not yet tapped the full information available in the finer grained HMM clusters. Looking at the upper bound, which is computed as the optimal selection of HMM states given knowledge of true argument boundaries, there is still some room for improvement.

6.2 Integrating into Online Latent Classifier

Next, we use this bottom-up argument identification system to constrain the argument search in our latent classifier training. During training, we restrict the set of possible argument structures (H_i in Algorithm 3) such that only those structures that agree with the HMM argument identification are considered, and the best labeling from the Superset of labels is selected for this structure. If we use the arguments identified via HMM argument identification to essentially fix the argument structure during training, the problem remaining for the learner is to select the predicate from among the non-argument content words in the sentence, while also identifying the labeling that is most consistent with the identified arguments and expectations from previous training.

Table 7 shows that once we add the HMM bottom-up argument identification to the Superset feedback scheme, the argument and predicate performance increases greatly (due to accuracy of the HMM argument identification). Note in Table 7 that bottom-up HMM argument identification is strong (0.88 F1 compared to 0.93 when

trained with true arguments), and that this effective argument-identification in turn permits strong performance on verb identification. Thus our procedure for tagging some HMM classes as argument (noun) classes based on a seed set of concrete nouns, combined with ambiguous Superset semantic feedback that does not indicate the number or identity of semantic arguments, yields enough information to begin learning to identify predicates (verbs) in input sentences.

Next, looking at the final role classification performance of the Superset+ argument constraint training schemes in Table 7, we see that Role F1 increases over both straight Superset and unordered Set of Labels feedback schemes. This increase is most dramatic for the more common A0 and A1 roles.

This represents one possible implementation of the structure-mapping procedure for early syntactic bootstrapping. If we assume the learner can learn some nouns with no guidance from syntactic knowledge (represented by our seed nouns), that noun knowledge can be combined with distributional learning (represented by our HMM parser) to tag some word-classes as noun classes. Representing each sentence as containing some number of these nouns (HMM argument identification) then permits the Latent BabySRL to begin learning to assign semantic roles to those nouns in sentences given highly ambiguous feedback, and also to use that ambiguous semantic feedback, combined with the constraints provided by the set of identified nouns in the sentence, to improve the latent syntactic representation, beginning to identify verbs in sentences.

This latent training method with ambiguous feedback works because it is seeking consistency in the features of the structures it sees. At the start of training, or when encountering a novel sentence with features not seen before, the latent inference will essentially choose a structure and labeling at random (since all structures will have the same score of 0, and ties are broken randomly). From this random labeling the classifier will increase connection strengths between lexical and structural features in the input sentence, and the (at first randomly) selected semantic role labels. Assuming that some number of random or quasi-random predictions are initially made, the learner can only improve if some feature weights increase above the others and begin to dominate predictions, both in the latent structure classifier and in the linked SRL classifier. This dominance can emerge only if there are structural features of sentences that frequently co-occur with frequent semantic roles.

Thus, the assignment of A0 and A1 roles can be learned by this latent SRL learner despite superset feedback, both because of the frequency of these two roles in the training data and their consistent co-occurrence with simple sentence-structure features that make use of the bottom-up information provided by the HMM argument identification. If “She likes yellow flowers.” is encountered early during latent training, the feedback may be the superset {A0, A1, A4, AM-LOC}, where the true labels A0 and A1 are present along with two other random labels. With accurate identification of ‘she’ and ‘flowers’ as arguments via the HMM bottom-up argument identification system, the learner will choose among only those role labelings that use two of the four roles. Given a large number of different sentences such as “She kicks the ball” (true labels are A0, A1), “She writes in her book” (A0, A2), and “She sleeps” (A0), the most consistent labeling amongst the true and random labelings

provided by Superset feedback will be that both 'she' and the first of two nouns are more likely to be labeled as A0. This consistent labeling is then propogated through the learner's weights, and used for future predictions and learning. Thus, even superset feedback can be informative given bottom-up information about the *nouns* in the sentence, because frequent nouns and argument patterns (e.g., first of two nouns) consistently co-occur with frequent roles (e.g., A0). Without the identified arguments, the chance of randomly assigning the correct arguments and roles decreases dramatically; as a result, the likelihood of encountering the correct interpretation often enough for it to dominate disappears.

7 Conclusion

We began with two problems for accounts of language acquisition: The sequences of words that make up the input sentences constitute highly ambiguous evidence for syntactic structure, and the situations in the world that accompany the input sentences constitute highly ambiguous evidence for sentence meaning. These two problems have led to 'bootstrapping' approaches to language acquisition, in which some set of built-in representational or architectural constraints on the language-learning system permit the learner to infer one type of structure from knowledge of another. Via semantic bootstrapping [70, 71], the learner uses independent knowledge of word and sentence meaning to identify the covert syntactic structures of sentences. Via syntactic bootstrapping [35, 40, 54, 65], the learner uses independent partial knowledge of syntactic structures to determine sentence meaning. These views are sometimes described as competing accounts, but in fact they share many assumptions, crucially including the assumption that the learner begins with some constraints on the possible links between syntax and semantics. In the present work we tried to incorporate key intuitions of both semantic and syntactic bootstrapping accounts to jointly address the two ambiguity problems with which we began.

To do so, we created a system within which we could manipulate the provision of partially-reliable syntactic and semantic information sources during language acquisition. We trained a semantic role classifier jointly with a simplified latent syntactic structure classifier, with learning based on (varyingly ambiguous) semantic feedback and simple linguistic constraints. This Latent BabySRL, sketched in Fig. 1, began by using an HMM to cluster unlabeled word-forms by part of speech. This clustering was based on distributional information recoverable from input word sequences, and was constrained by an initial division into content and function words, and by prior biases regarding the sparsity of word classes. This step represents the assumption that infant learners, even before they understand the meanings of words or sentences, gather statistics about how words are distributed in the linguistic input (e.g., [42, 58, 76]), and discriminate content from function words based on their distinct phonological properties [77, 78]. It is well established in previous work that considerable information about grammatical category similarity can be obtained by the kind of sequential distributional analysis that an HMM

undertakes. We assume that other learning architectures that are sensitive to the sequential statistics of the input would produce similar results; this would include a simple recurrent network that learns a category structure in its hidden units to predict the next word in input sentences (e.g., [14, 29]).

With this previous distributional learning in hand, the Latent BabySRL attempted to jointly learn a latent structure for identifying arguments (nouns) and a predicate (verb) in input sentences, and to predict the roles of the identified arguments relative to the identified predicate. The only information sources for this joint learning task were the semantic-role feedback (ranging from full 'gold standard' feedback to highly ambiguous superset feedback) provided to the semantic-role classifier, the representational constraints on the two classifiers (their feature sets), and the way in which the predictions of the latent structure classifier were used to generate input features for the semantic role classifier. These constraints represent simple but substantive constraints on the links between syntax and semantics. First, the semantic role classifier predicts a semantic role for all and only the nouns it finds in the input sentence. This represents a simple built-in link between syntax and semantics, and a key assumption of the structure-mapping view: the learner assumes each noun is an argument of some predicate term. Second, the latent structure classifier and the semantic-role classifier are equipped with both lexical features (the words in the sentence) and more abstract structural features that permit them to generalize beyond particular words. These abstract features include the predicted HMM clusters of the nouns and verb identified in the sentence, and also simple sentence-structure relational features that can be derived from the identified sequence of nouns and verb, features such as "1st of 2 nouns" and "preverbal noun". Crucially, the specific content and the connection weights of these simple abstract structural features are not provided to the model as hand-coded features of input sentences; such a choice would model a learner that (somehow) begins with accurate identification of nouns and verbs. Instead, the specific syntactic and semantic knowledge that develops in the system arises from the kinds of features the classifiers can represent, and the way in which the model is set up to use them to identify latent structures and in turn to predict semantic roles. Thus we model a learner that begins with substantive constraints on links between syntax and semantics, but without being informed of which words are nouns and which are verbs.

When trained with very informative semantic-role feedback, the Latent BabySRL implements a simple form of semantic bootstrapping. The provision of complete semantic role feedback represents the assumption that the child knows the meaning of the content words in the sentence, and can generate an interpretation of the input sentence based on observing the accompanying scene. Given such veridical semantic feedback, the Latent BabySRL can straightforwardly identify the noun arguments in the sentence (they are the ones that play semantic roles such as agent or patient), but can also learn to identify the verb, by learning that the identity and position of the verb are useful predictor of semantic roles in English sentences (e.g., preverbal nouns tend to be agents).

When trained with highly ambiguous semantic feedback, the Latent BabySRL still learned to identify arguments and predicates, and to use that inferred syntactic structure to assign semantic roles, but only if the system was 'primed' with knowledge of a small set of concrete nouns. The superset feedback described in Sect. 5.1 made possible many interpretations of each input sentence (including the true one); this feedback scheme provided no information about the number of arguments in each sentence, or which word in the sentence should be aligned with each semantic role. We implemented a procedure whereby a set of concrete seed nouns was used to automatically tag some HMM clusters as noun clusters. This bottom-up argument identification system then constrained the argument search in the latent structure classifier training, as described in Sect. 6.2. Representing each input sentence as containing some number of nouns guided the learner's assignment of meaning to input sentences; this in turn permitted the Latent BabySRL to improve its representation of input sentences (including learning to identify the verb), and therefore to further improve its semantic-role classification.

This process represents one straightforward implementation of the structure-mapping account of the origin of syntactic bootstrapping. A skeletal sentence structure, grounded in a set of concrete nouns, provides a preliminary estimate of the number and identity of the noun arguments in the sentence, which in turn permits further semantic and syntactic learning. The Latent BabySRL's dramatic failure to learn when provided with superset feedback without this bottom-up information about the number of noun arguments in the sentence suggests that argument-number information, which in principle could be derived from lucky observations of informative scenes (as in the full-feedback version), or from partial knowledge of syntax grounded in a set of nouns, was crucial to permitting the system to learn.

One might ask which of these two settings of our model is closer to the typical state of the human learner. Should we assume the semantic bootstrapping setting is typical – that the child often knows the meanings of the content words in sentences, and can divine the sentence's meaning from observation of scenes? Or should we assume that the syntactic bootstrapping setting is typical, particularly at early points in acquisition – that the child needs guidance from the sentence itself to determine the abstract relational meanings of verbs, and of sentences? Some would argue that even toddlers can often determine the speaker's intended message in contexts of face-to-face interaction, reading the speaker's intention in a shared interactional goal space (e.g., [71, 83]). Others, including the present authors, would argue that the abstract relational meanings of verbs and sentences cannot routinely be determined from event observation without linguistic guidance (e.g., [32, 40, 75]). The present computational experiments contribute to this conversation by making explicit one way in which partial representations of the structure of sentences, derived with the aid of no semantic information beyond the meanings of a few concrete nouns, to guide early verb learning and sentence interpretation.

Acknowledgements We wish to thank Yael Gertner for insightful discussion that led up to this work as well as the various annotators who helped create the semantically tagged data. This research is supported by NSF grant BCS-0620257 and NIH grant R01-HD054448.

References

1. Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes, 25*(1), 50–93.
2. Alishahi, A., & Stevenson, S. (2012). Gradual acquisition of verb selectional prefences in a bayesian model. In A. Villavicencio, A. Alishahi, T. Poibeau, & A. Korhonen (Eds.), *Cognitive aspects of computational language acquisition*. New York: Springer.
3. Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.
4. Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.
5. Bloom, B. H. (1970). Space/time trade-offs in Hash coding with allowable errors. *Communications of the ACM, 13*(7), 422–426.
6. Bloom, L. (1973). *One word at a time: The use of single-word utterances before syntax*. The Hague: Mouton.
7. Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*(5), 752–793.
8. Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*, 31–44.
9. Brill, E. (1997). Unsupervised learning of disambiguation rules for part of speech tagging. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Eds.), *Natural language processing using very large corpora*. Dordrecht: Kluwer Academic Press.
10. Brown, R. (1973). *A first language*. Cambridge: Harvard University Press.
11. Brown, P., Pietra, V. D., deSouza, P., Lai, J., & Mercer, R. (1992). Class-based n-gram models of natural language. *Computational Linguistics, 18*(4), 467–479.
12. Carreras, X., & Màrquez, L. (2004). Introduction to the CoNLL-2004 shared tasks: Semantic role labeling. In *Proceedings of CoNLL-2004* (pp. 89–97), Boston.
13. Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor.
14. Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review, 113*(2), 234–272.
15. Chang, M., Goldwasser, D., Roth, D., & Srikumar, V. (2010). Discriminative learning over constrained latent representations. In *Proceedings of the Annual Meeting of the North American Association of Computational Linguistics (NAACL)*, Los Angeles.
16. Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Providence.
17. Cherry, C., & Quirk, C. (2008). Discriminative, syntactic language modeling through latent svms. In *Proceedings of the Eighth Conference of AMTA*, Honolulu.
18. Clark, E. V. (1978). Awwareness of language: Some evidence from what children say and do. In R. J. A. Sinclair & W. Levelt (Eds.), *The child's conception of language*. Berlin: Springer.
19. Clark, E. V. (1990). Speaker perspective in language acquisition. *Linguistics, 28*, 1201–1220.
20. Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, Philadelphia.
21. Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2008). Baby srl: Modeling early language acquisition. In *Proceedings of the Annual Conference on Computational Natural Language Learning (CoNLL)*, Manchester.
22. Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2009). Minimally supervised model of early language acquisition. In *Proceedings of the Annual Conference on Computational Natural Language Learning (CoNLL)*, Boulder.
23. Connor, M., Gertner, Y., Fisher, C., & Roth, D. (2010). Starting from scratch in semantic role labeling. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*, Uppsala.

24. Connor, M., Fisher, C., & Roth, D. (2011). Online latent structure training for language acquisition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona.
25. Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125–127.
26. Demetras, M., Post, K., & Snow, C. (1986). Feedback to first-language learners. *Journal of Child Language, 13*, 275–292.
27. Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of english. *Language & Speech, 49*, 137–174.
28. Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*, 547–619.
29. Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.
30. Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.
31. Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) Anchorage*, Alaska.
32. Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretation of sentences. *Cognitive Psychology, 31*, 41–81.
33. Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child Development, 67*, 3192–3218.
34. Fisher, C., Gleitman, H., & Gleitman, L. (1989). On the semantic content of subcategorization frames. *Cognitive Psychology, 23*, 331–392.
35. Fisher, C., Gertner, Y., Scott, R., & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 143–149.
36. Gao, J., & Johnson, M. (2008). A comparison of bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of EMNLP-2008* (pp. 344–352), Honolulu.
37. Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek & R. Golinkoff (Eds.), *Action meets word: How children learn verbs* (pp. 544–564). Oxford/New York: Oxford University Press.
38. Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science, 17*, 684–691.
39. Gildea, D., & Palmer, M. (2002). The necessity of parsing for predicate argument recognition. In *ACL* (pp. 239–246), Philadelphia.
40. Gillette, J., Gleitman, H., Gleitman, L. R., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*, 135–176.
41. Goldwater, S., & Griffiths, T. (2007). A fully bayesian approach to unsupervised part-of-speech tagging. In *ACL* (pp. 744–751), Prague.
42. Gomez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.
43. Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. In *Proceedings of HTL-NAACL*, New York.
44. Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, Boulder.
45. Harris, Z. (1951). *Methods in structural linguistics*. Chicago: Chicago University Press.
46. Hochmann, J., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition, 115*, 444–457.
47. Huang, F., & Yates, A. (2009). Distributional representations for handling sparsity in supervised sequence-labeling. In *ACL*, Singapore.
48. Johnson, M. (2007). Why doesn't EM find good HMM POS-taggers? In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL* (pp. 296–305), Prague.

49. Johnson, M., Demuth, K., Frank, M. C., & Jones, B. (2010). Synergies in learning words and their meanings. In *Neural Information Processing Systems, 23*, Vancouver.
50. Kazama, J., & Torisawa, K. (2007). A new perceptron algorithm for sequence labeling with non-local features. In *Proceedings of the 2007 Joint Conference of EMNLP-CoNLL* (pp. 315–324), Prague.
51. Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*, 349–364.
52. Kingsbury, P., & Palmer, M. (2002). From Treebank to PropBank. In *Proceedings of LREC-2002*, Spain.
53. Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the Association for Computational Linguistics (ACL)*, Barcelona.
54. Landau, B., & Gleitman, L. (1985). *Language and experience*. Cambridge: Harvard University Press.
55. Levin, B., & Rappaport-Hovav, M. (2005). *Argument realization. Research surveys in linguistics series*. Cambridge: Cambridge University Press.
56. MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah: Lawrence Elrbaum Associates.
57. Marcus, M. P., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.
58. Marcus., G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77–80.
59. Màrquez, L., Carreras, X., Litkowski, K., & Stevenson, S. (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics, 34*, 145–159.
60. Meilă, M. (2002). *Comparing clusterings* (Tech. Rep. 418). University of Washington Statistics Department.
61. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–312.
62. Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91–117.
63. Mintz, T., Newport, E., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393–424.
64. Monaghan, P., Chater, N., & Christiansen, M. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96*, 143–182.
65. Naigles, L. R. (1990). Children use syntax to learn verb meanings. *Journal of Child Language, 17*, 357–374.
66. Nappa, R., Wessel, A., McEldoon, K., Gleitman, L., & Trueswell, J. (2009). Use of speaker's gaze and syntax in verb learning. *Language Learning and Development, 5*, 203–234.
67. Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71–106.
68. Parisien, C., & Stevenson, S. (2010). Learning verb alternations in a usage-based bayesian model. In *Proceedings of the 32nd annual meeting of the Cognitive Science Society*, Portland.
69. Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language, 37*, 607–642.
70. Pinker, S. (1984). *Language learnability and language development*. Cambridge: Harvard University Press.
71. Pinker, S. (1989). *Learnability and cognition*. Cambridge: MIT Press.
72. Punyakanok, V., Roth, D., & Yih, W. (2008). The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics, 34*(2), 257–287.
73. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–285.

74. Ravi, S., & Knight, K. (2009). Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conferenceof the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, Singapore.
75. Rispoli, M. (1989). Encounters with japanese verbs: Caregiver sentences and the categorization of transitive and intransitive action verbs. *First Language, 9*, 57–80.
76. Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 906–914.
77. Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*(01), 169–201.
78. Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*(2), B11–B21.
79. Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558–1568.
80. Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon*. Cambridge: MIT Press.
81. Solan, Z., Horn, D., Ruppin, E., & Edelman, S. (2005). Unsupervised learning of natural languages. *Proceedings of the National Academy of Science, 102*, 11629–11634.
82. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Manchester.
83. Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.
84. Toutanova, K., & Johnson, M. (2007). A bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Proceedings of NIPS*, Vancouver.
85. Waterfall, H., Sandbank, B., Onnis, L., & Edelman, S. (2010). An empirical generative framework for computational modeling of language acquisition. *Journal of Child Language, 37*, 671–703.
86. Yang, C. (2011). A statistical test for grammar. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, Portland.
87. Yu, C., & Joachims, T. (2009). Learning structural svms with latent variables. In *ICML*, Montreal.
88. Yuan, S., Fisher, C., & Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development, 83*, 1382–1399.

Gradual Acquisition of Verb Selectional Preferences in a Bayesian Model

Afra Alishahi and Suzanne Stevenson

Abstract We present a cognitive model of inducing verb selectional preferences from individual verb usages. The selectional preferences for each verb argument are represented as a probability distribution over the set of semantic properties that the argument can possess—a *semantic profile*. The semantic profiles yield verb-specific conceptualizations of the arguments associated with a syntactic position. The proposed model can learn appropriate verb profiles from a small set of noisy training data, and can use them in simulating human plausibility judgments and analyzing implicit object alternation.

1 Introduction

Many verbs show strong preferences concerning the semantic properties of their arguments. For example, *eating food* and *drinking water* are acceptable, whereas **eating water* and **drinking food* are normally not. Learning verb selectional preferences is an important aspect of human language acquisition, and the acquired preferences have been shown to guide children's and adults' expectations about missing or upcoming arguments in language learning and comprehension (e.g., [13, 24]).

In the earlier theories of semantics, verbs are assumed to impose limitations or constraints on the applicability of potential arguments filling a particular role

A. Alishahi (✉)
Department of Communication and Information Studies, Tilburg University, Tilburg, The Netherlands
e-mail: a.alishahi@uvt.nl

S. Stevenson
Department of Computer Science, University of Toronto, Toronto, ON, Canada
e-mail: suzanne@cs.toronto.edu

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4_11,

(e.g., [15, 17]). However, this view has been challenged by an alternative approach proposed by Resnik [29], in which predicates (particularly verbs) show preferences towards certain arguments, as opposed to constraining the set of arguments that they can take. Resnik [29] introduced a statistical approach to automatic induction of verb selectional preferences from a corpus. In this framework, a semantic class hierarchy for words is used, together with statistical tools, to induce a verb's selectional preferences for a particular argument position in the form of a probability distribution over all the classes that can occur in that position. Resnik's model was proposed as a model of human learning of selectional preferences that made minimal representational assumptions; it showed how such preferences could be acquired from usage data and an existing conceptual hierarchy.

The computational study of learning verb selectional preferences is heavily influenced by the model of Resink [29]. However, his and later computational models (see Sect. 1.2) have properties that do not match with certain cognitive plausibility criteria for a child language acquisition model. All of these models use the training data in "batch mode", and most of them use information-theoretic measures that rely on total counts from a corpus. Therefore, it is not clear how the representation of selectional preferences could be updated incrementally in these models as the person receives more data. Moreover, the assumption that children have access to a full hierarchical representation of semantic classes may be too strict.

In this paper, we propose a cognitive model of the representation and acquisition of verb selectional preferences which is more plausible in the context of child language acquisition. We present an incremental Bayesian model for inducing selectional preferences from usage data. In this model, the selectional preferences of a verb are represented as a probability distribution over the *semantic properties* of an argument, and are evolved over time as the model observes more usages of each verb. In a series of experiments, we show that our model can form intuitive selectional preferences for a range of verbs, and make appropriate generalizations over the observed properties, which can be used in simulating human plausibility judgments.

1.1 Verb Selectional Preferences

Selectional preferences, or constraints, are viewed as limitations on the applicability of natural language predicates to arguments. In their semantic theory, Katz and Fodor [17] characterized selectional constraints as restrictions in terms of the defining features of the arguments: they outlined a decompositional theory of word meaning in which lexical entries specified the features applicable to a particular lexical item. For words that denote predicates, Katz and Fodor proposed that the arguments in their lexical entries be annotated with restrictions identifying the necessary and sufficient conditions that a semantically acceptable argument must meet. Such conditions were represented as Boolean functions of semantic

features, such as HUMAN or HIGHER ANIMAL for the subject of the verb *hit*, and PHYSICAL OBJECT for its object. Jackendoff's [15] lexical theory, on the other hand, situates selectional constraints as information appearing in the context of a rich representation of the predicate's meaning, such as the annotation LIQUID appearing as a constraint on one argument of the verb *drink*. Selectional constraints were also explicitly integrated into grammar, as in Generative Lexicon Theory [9, 27]. However, identifying restrictions that are both necessary and sufficient, and choosing the primitives themselves, is viewed by many to be an insurmountable problem.

Resnik [28] instead emphasized the view of the restrictions a verb places on its arguments as selectional preferences, and proposed a new approach to their representation and learning, which was followed by many in the computational linguistics community. In this approach, the knowledge of words, or concepts, is represented as a pre-defined semantic class hierarchy, and statistical tools are used to learn selectional preferences from examples in a corpus. As opposed to a Boolean interpretation of selectional constraints, here the selectional preferences are viewed as probability distributions over various semantic classes. For example, the preferred objects of *eat* are represented not as the black-and-white class FOOD but rather as a gray probability distribution over all nouns or various classes thereof.

Many theories of lexical acquisition make use of selectional constraints [13, 26]. Gleitman and Gillette [13] show that selectional constraints provide adult subjects with significant constraints on the possible meanings of unknown verbs: the subjects identified a verb 80 % of the time if they were given the syntactic frame of the verb together with the nouns that appear as the verb arguments; however, the syntactic frame alone or the noun arguments alone (without specifying their position) did not help subjects to identify the verb half the time. This shows that the semantic properties of the verb arguments (or verb selectional preferences) are more informative than simply the semantic associations between a verb and a group of nouns, or the syntactic properties of the verb. Moreover, selectional preferences play an important role in many aspects of language processing: they influence the syntactic structure of a sentence, especially in the face of ambiguity; they affect selecting the likely word in a sequence of speech signals; and they can be drawn on for the task of word sense disambiguation. An explicit model of the process by which the acquisition of selectional preferences takes place can shed light on the plausible representations and their effect on the relevant language tasks.

1.2 Related Computational Models

Two central questions for the automated treatment of selectional preferences are: what *representation* to use, and how to *induce* preferences from available data. A variety of computational models of verb selectional preferences have been proposed, which use different statistical models to induce the preferences of each verb from corpus data. Most of these models, however, use the same representation

for verb selectional preferences: the preference can be thought of as a mapping, with respect to an argument position for a verb, of each semantic class to a real number [19]. The induction of a verb's preferences is, therefore, modeled as using a set of training data to estimate that number.

Resnik [28] is the first to model the problem of induction of selectional preferences using a a pre-existing semantic class hierarchy, WordNet [23]. He defines the selectional preference strength of a verb for a particular argument as the divergence between two probability distributions: the prior probabilities of the classes in that argument position (e.g., direct object), and the posterior probabilities of the classes in that position given that verb. The selectional association of a verb with a class is also defined as the contribution of that class to the total selectional preference strength. For example, *eat* would be expected to have a reasonably strong selectional preference strength, with food items having high selectional association and non-food items a very low selectional association. Resnik estimates the prior and posterior probabilities based on the frequencies of each verb and its relevant argument in a corpus.

Following [28, 29], a number of methods were presented that make use of WordNet and a text corpus, together with a variety of statistical models, to induce selectional preferences. Li and Abe [18] model selectional preferences of a verb (for an argument position) as a set of nodes in WordNet with a probability distribution over them. They use the Minimum Description Length (MDL) principle to find the best set for each verb and argument based on the usages of that verb in the training data. Clark and Weir [7] also find an appropriate set of concept nodes to represent the selectional preferences for a verb, but do so using a χ^2 test over corpus frequencies mapped to concepts to determine when to generalize from a node to its parent. Ciaramita and Johnson [6] use a Bayesian network with the same topology as WordNet to estimate the probability distribution of the relevant set of nodes in the hierarchy. Abney and Light [1] use a different representational approach: they train a separate hidden Markov model for each verb, and the selectional preference is represented as a probability distribution over words instead of semantic classes.

In contrast to the class-based methods above, Erk [11] proposes a similarity equation-based model that does not rely on a hierarchical representation of semantic classes. Instead, her model estimates the selectional preference of an argument position for a possible head word as a frequency-weighted sum of the similarities between that word and the observed head words for that argument position in a corpus. The similarity between the potential head word and each previously-observed head word is computed based on a corpus-based semantic similarity metric. Zapirain et al. [32] use a similar approach for automatically generating selectional preferences from a corpus using a second-order distributional similarity measure, which they use in semantic role classification. Such similarity-based models generally outperform the class-based approaches for many tasks, but are unable to form an abstract representation of selectional preferences.

It is not easy to evaluate the acquired selectional preferences on their own, since there is no "gold standard" set of examples against which to compare the outcome of a method [19]. The existing models of verb selectional preference have

been evaluated through a wide range of computational linguistic tasks, including word sense disambiguation [1, 6, 22, 30], PP-attachment disambiguation [18]; a pseudo-disambiguation task of choosing the best verb-argument pair [7], and semantic role labelling [11]. Resnik [29] also evaluated his method through two other means that are more interesting from a human language acquisition point of view: the simulation of verb-argument plausibility judgements elicited from adult subjects, and an analysis of whether implicit verb arguments—those that are not syntactically realized—are those that are strongly semantically constrained. We refer to the simulation of human plausibility judgments in our experimental results.

1.3 Our Approach

In previous work [3], we have proposed a usage-based model of early verb learning that uses Bayesian clustering and prediction to model language acquisition and use. Individual verb usages are incrementally grouped to form emergent classes of linguistic constructions that share syntactic and semantic properties. We have shown that our Bayesian model can incrementally acquire a general conception of the semantic roles of predicates based only on exposure to individual verb usages [4]. The model forms probabilistic associations between the semantic properties of arguments, their syntactic position, and the semantic primitives of verbs. Our previous experiments demonstrated that, initially, this probability distribution for an argument position yields verb-specific conceptualizations of the role associated with that position. As the model is exposed to more input, the verb-based roles gradually transform into more abstract representations that reflect the general properties of arguments across the observed verbs. See also [8] in this volume for an alternative approach to the acquisition of semantic roles.

In this paper, we present an extended version of our model that, in addition to learning general semantic roles for constructions, can use its verb-specific knowledge to predict intuitive selectional preferences for each verb argument position.[1] We propose a novel way of representing the selectional preferences of a verb as a *verb semantic profile*, or a probability distribution over the semantic properties of an argument for each verb. A verb semantic profile is predicted from both the verb-based and the construction-based knowledge that the model has learned through clustering, and reflects the properties of the arguments that are observed for that verb. Our proposed model makes appropriate generalizations over the observed properties, and captures expectations about previously unseen arguments.

As in other work on selectional preferences, the semantic representation of arguments in our model is based on a standard lexical ontology [WordNet; 23].

[1] This paper is an updated and extended version of preliminary work on this approach presented in [2].

Verb Usage: *We entered the room.*	
Extracted Frame	
head verb	*enter*
semantic primitives of verb:	⟨*register,record,enter,put down,save,preserve,keep,hold on, have,have got,hold,be,* ... ⟩
number of arguments:	2
syntactic pattern:	arg1 *verb* arg2
argument 1:	*we*
properties of argument 1:	⟨*organism,being,living thing,animate thing,object,physical object,entity,causal agent,* ... ⟩
argument 2:	*room*
properties of argument 2:	⟨*area,structure,construction,object,physical object,entity, whole,unit,position,spatial relation,*... ⟩

Fig. 1 An input sentence and its corresponding frame

In contrast to other work, however, each argument contributes to the semantic profile of the verb through a (potentially large) set of semantic properties instead of its membership in a class in the hierarchy. It should be emphasized that our model does not require knowledge of the hierarchical structure of the WordNet concepts. That is, the model is able to generalize knowledge of semantic classes without requiring an explicit class structure; all that is required for generalization behaviour is that some properties are more general (i.e., shared by more words) than others. In other aspects, the particular semantic properties are not fundamental to the working of the model, and could in the future be replaced with another resource that is deemed more appropriate to child language acquisition.

This approach allows us the computational advantage of making use of an available resource, while avoiding ad hoc cognitive assumptions about the representation of a conceptual hierarchy. Moreover, due to our novel representation of a semantic profile, the model can induce and use selectional preferences using a relatively small set of training data.

2 A Computational Model of Learning Verb Selectional Preferences

Our model learns the set of *argument structure frames* for each verb, and their grouping across verbs into *constructions*. An argument structure frame is a set of features of a verb usage that are both syntactic (the number of arguments, the syntactic pattern of the usage) and semantic (the semantic properties of the verb, the semantic properties of each argument). The syntactic pattern indicates the word order of the verb and arguments. Figure 1 shows an example of a verb usage and the corresponding argument structure frame.

A construction is a grouping of individual frames which probabilistically share syntactic and semantic features, and form probabilistic associations across verb

semantic properties, argument semantic properties, and the syntactic pattern. These groupings typically correspond to general constructions in the language such as transitive, intransitive, and ditransitive.

For each verb, the model associates an argument position with a semantic profile, which is a probability distribution over a set of semantic properties. In doing so, the model uses the knowledge that it has learned for that verb, as well as the grouping of frames for that verb into constructions.

The model formalization is presented in the following sections. We review basic properties of the model, i.e. the clustering of argument structure frames into constructions (Sect. 2.1) and the estimation of the probabilities of semantic properties (Sect. 2.2) from [3, 4].[2] Next we describe the extensions that give the model its ability to make verb-based predictions: in Sect. 2.3, a novel approach for estimating a semantic profile for each argument position of a verb is presented, and in Sect. 2.4 a new criteria for measuring the compatibility between a verb and an argument is proposed.

2.1 Learning as Bayesian Clustering

Each argument structure frame for an observed verb usage is input to an incremental Bayesian clustering process. This process groups the new frame together with an existing group of frames—a construction—that probabilistically has the most similar semantic and syntactic properties to it. If no construction has sufficiently high probability for the new frame, then a new construction is created for it. We use the probabilistic model of [3, 4] for learning constructions, which is itself an adaptation of a Bayesian model of human categorization proposed by Anderson [5]. It is important to note that the categories (i.e., constructions) are not predefined, but rather are created according to the patterns of similarity over observed frames.

Grouping a frame F with other frames participating in construction k is formulated as finding the k with the maximum probability given F:

$$\textbf{BestConstruction}(F) = \underset{k}{\text{argmax}}\ P(k|F) \tag{1}$$

where k ranges over the indices of all constructions, with index 0 representing creation of a new construction.

Using Bayes rule, and dropping $P(F)$ which is constant for all k:

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k) \tag{2}$$

[2] Please refer to [3, 4] for more details about the representation framework and the learning mechanisms, and more in-depth discussions about the motivation behind each step.

The prior probability, $P(k)$, indicates the degree of entrenchment of construction k, and is given by the relative frequency of its frames over all observed frames. The posterior probability of a frame F is expressed in terms of the individual probabilities of its features, which (under an assumption of independence) yields a simple product of feature probabilities:

$$P(F|k) = \prod_{i \in FrameFeatures} P_i(j|k) \tag{3}$$

where j is the value of the ith feature of F, and $P_i(j|k)$ is the probability of displaying value j on feature i within construction k (determined by considering the values of feature i across all frames in k). If no existing construction has a sufficiently high similarity to the features of the current frame F to be clustered, the new construction, $k = 0$, will have the highest probability, and the current frame F will trigger creation of a new cluster.

Given the focus here on semantic profiles, we next focus on the calculation of the probabilities of semantic properties.

2.2 Probabilities of Semantic Properties

The probability in Eq. (3) of value j for feature i in construction k is estimated using a smoothed version of this maximum likelihood formula:

$$P_i(j|k) = \frac{\text{count}_i^k(j)}{n_k} \tag{4}$$

where n_k is the number of frames participating in construction k, and $\text{count}_i^k(j)$ is the number of those with value j for feature i.

For most features, $\text{count}_i^k(j)$ is calculated by simply counting those members of construction k whose value for feature i exactly matches j. However, recall that for the semantics of a word, the value of this feature is a (typically, large) set of properties. Counting only the number of exact matches between sets of such properties is too strict, since even highly similar words very rarely have the exact same set of properties. We instead use the following Jaccard similarity score to measure the overlap between the set of semantic properties, S_i^F, of a particular argument i in the frame F to be clustered, and the set of semantic properties, S_i^k, of the same argument in a member frame k of a construction:

$$\text{sem\_score}(S_i^F, S_i^k) = \frac{|S_i^F \cap S_i^k|}{|S_i^F \cup S_i^k|} \tag{5}$$

The conditional probability of a set of semantic properties ($P_i(S_i^F|k)$) is calculated as in Eq. (4), but $\text{count}_i^k(j)$ is estimated as the normalized sum of

the sem_scores for the new frame F and every member of construction k. The normalization factor is calculated as the sum of the similarity of every word in our lexicon to the frames in construction k.

2.3 Predicting Semantic Profiles for Verbs

The formula in Eq. (4), $P_i(j|k)$, is used to calculate the probability of the semantics of an argument for a given position across all usages in a construction. This can be used to provide a general picture of the semantic role for that position across all verbs in the construction. To see specifically what arguments a particular verb prefers in that position, we need a different probability formula restricted to that verb, across all its usages.

We represent the selectional preferences of a verb for an argument position as a verb semantic profile, which is a probability distribution over all the semantic properties. To predict the profile of a verb v for an argument position *arg*, we need to estimate the probability of each semantic property j separately.

$$\begin{aligned} P_{arg}(j|v) &= \sum_k P_{arg}(j,k|v) \\ &\propto \sum_k P(k,v)\, P_{arg}(j|k,v) \end{aligned} \tag{6}$$

Here, j ranges over all the possible semantic properties that an argument can have, and k ranges over all constructions.

The prior probability of having verb v in construction k, or $P(k,v)$, is mainly determined by the frequency with which v participates in k, or f_k^v. However, due to sparse and noisy data, individual verb usages are not entirely reliable. On the other hand, the items within a well-entrenched construction are more likely to reflect a reasonably confident usage of the verb. We thus want to include the influence of the degree of entrenchment of each construction in the prior probability. To determine the relative entrenchment of a construction k among all the constructions, we calculate its weight w_k by taking its number of frames n_k over the total number of frames:

$$w_k = \frac{n_k}{\sum_{k'} n_{k'}} \tag{7}$$

The prior probability $P(k,v)$ is then calculated by:

$$P(k,v) = \frac{w_k \times f_k^v}{\sum_{k'} w_{k'} \times f_{k'}^v} \tag{8}$$

The posterior probability $P_{arg}(j|k,v)$ is calculated analogously to $P_i(j|k)$ in Eq. (4), but limiting the count of matching features to those frames in k that

contain v:

$$P_{arg}(j|k,v) = \frac{\text{verb\_count}^{k}_{arg}(j,v)}{n_{kv}} \tag{9}$$

where n_{kv} is the number of frames for v participating in construction k, and $\text{verb\_count}^{k}_{arg}(j,v)$ is the number of those with semantic property j for argument *arg*. We use a smoothed version of the above formula, where the relative frequency of each property j among all nouns is used as the smoothing factor.

2.4 Verb-Argument Compatibility

To estimate verb argument plausibility, we need a measure of compatibility of a particular noun n for an argument position *arg* of some verb v. That is, we need to estimate how much the semantic properties of n conform to the acquired semantic profile of v for *arg*. We formulate the compatibility as the log of the conditional probability of observing n as an argument *arg* of v:

$$\text{compatibility}(v,n) = \log(P_{arg}(\text{prop}(n)|v)) \tag{10}$$

where $\text{prop}(n)$ is the set of the semantic properties for word n, and $P_{arg}(\text{prop}(n)|v)$ is estimated similarly to $P_{arg}(j|v)$ in Eq. (6). Since $\text{prop}(n)$ is a set of properties (as opposed to j in Eq. (6) being a single property), $\text{verb\_count}^{k}_{arg}$ in Eq. (9) should be modified as:

$$\text{verb\_count}^{k}_{arg}(\text{prop}(n),v) = \sum_{f \in k} \text{sem\_score}(\text{prop}(n), S^{f}_{arg}) \tag{11}$$

where f is a frame that belongs to construction k, and S^{f}_{arg} is the set of the semantic properties for argument *arg* of frame f.

3 Experimental Results

In the following sections, we first describe the training data for our model. Although our model determines selectional preferences for any argument position, we focus here on evaluation of verb preferences for the direct object position, as is typical in other computational models of selectional preferences. Next, we provide a qualitative analysis of our model through examination of the semantic profiles for a number of verbs, and show how the semantic profiles of verbs evolve over time. We then evaluate our model through simulating human judgments of verb-argument plausibility, following [29].

3.1 The Training Data

In earlier work [3, 4], we used a method to automatically generate training data with the same distributional properties as the input children receive. However, this relies on manually-compiled data about verbs and their argument structure frames from the CHILDES database [20]. To evaluate the new version of our model for the task of learning selectional preferences, we need a wide selection of verbs and their arguments that is impractical to compile by hand.

The training data for our experiments here are generated as follows. We use the Wall Street Journal portion of Penn Treebank [21]. We convert the treebank to a dependency format using the LTH Constituent-to-Dependency Conversion tool [16], from which we extract verb usages. For each verb usage in a sentence, we construct a frame by recording the lemmatized verb form, the number of the arguments for that verb, and the syntactic pattern of the verb usage (i.e., the position of the verb and the arguments).

We also record in the frame the semantic properties of the verb and each of the argument heads (each noun is also lemmatized). The semantic properties of words are taken from WordNet (version 2.0) as follows. In order to simulate understanding of the semantics of a word ranging from more specific to more general aspects, we use properties that reflect the nearer and more distant hypernyms of the word in WordNet. We extract all the hypernyms (ancestors) for all the senses of the word, and add all the words in the hypernym synsets to the list of the semantic properties. Figure 2 shows an example of the hypernyms for *dinner*, and its resulting set of semantic properties.[3]

3.2 Formation of Semantic Profiles for Verbs

We train our model on 30,000 frames extracted from WSJ (as described in the previous section). We use Eq. (6) to predict the semantic profile of the direct object position for a range of verbs. Some of these verbs, such as *pay* and *cause*, have strong selectional preferences, whereas others, such as *take* and *put*, can take a wide range of nouns as direct object (as confirmed by Resnik's [29] estimated strength of selectional preference for these verbs). Figure 3 displays the semantic profiles of three verbs: *pay*, *cause* and *put*. (Due to limited space, we only include the 30 properties that have the highest probability in each profile.)

Because we extract the semantic properties of words from WordNet which has a hierarchical structure, the properties that come from nodes in the higher levels of the hierarchy (such as *entity* and *abstraction*) appear as the semantic property for a very

[3] We do not remove alternate spellings of a term in WordNet; this will be seen in the profiles in the results section.

```
Sense 1
dinner
       => meal, repast
           => nutriment, nourishment, nutrition, sustenance,
              aliment, alimentation, victuals
               => food, nutrient
                   => substance, matter
                       => entity
Sense 2
dinner, dinner party
       => party
           => social gathering, social affair
               => gathering, assemblage
                   => social group
                       => group, grouping
```

dinner: {meal, repast, nutriment, nourishment, nutrition, substance, aliment, alimentation, victuals, food, nutrient, substance, matter, entity, party, social gathering, social affair, gathering, assemblage, social group, group, grouping}

Fig. 2 Semantic properties for *dinner* from WordNet

large set of words, whereas the properties that come from the leaves in the hierarchy are specific to a small set of words. Therefore, the general properties are more likely to be associated with a higher probability in the semantic profiles for most verbs. In fact, a closer look at the semantic profiles for less selective verbs such as *put* reveals that the top portion of the semantic profile for these verbs consists solely of such general properties that are shared among many words. However, this is not the case for the more restrictive verbs. The semantic profiles for *pay* and *cause* show that the specific properties that these verbs demand from their direct object appear amongst the highest-ranked properties, even though only a small set of words share these properties (e.g., *possession, transferred property, financial loss, cost, ...* for *pay*, and *human action, change, happening, occurrence, natural event, ...* for *cause*).

The examination of the semantic profiles for fairly frequent verbs in the training data shows that our model can use the verb usages to predict an appropriate semantic profile for each verb. When presented with a novel verb (for which no verb-based information is available), Eq. (6) predicts a semantic profile which reflects the relative frequencies of the semantic properties among all words (due to the smoothing factor added to Eq. (9)), modulated by the prior probability of each construction. The predicted profile is displayed in Fig. 4. It shows similarities with the profile for *put* in Fig. 3, but the general properties in this profile have an even higher probability. Since the profile for the novel verb is predicted in the absence of any evidence (i.e., verb usage) in the training data, we later use it as the base for estimating other verbs' strength of selectional preference.

	pay
(0.017)	abstraction
(0.014)	possession
(0.013)	entity
(0.012)	object
(0.012)	physical object
(0.012)	destination
(0.011)	relation
(0.011)	goal
(0.010)	transferred property
(0.010)	transferred possession
(0.010)	loss
(0.010)	financial loss
(0.010)	cost
(0.010)	outlay
(0.010)	outgo
(0.010)	expenditure
(0.010)	communication
(0.010)	social relation
(0.009)	measure
(0.008)	act
(0.008)	cause
(0.008)	instrument
(0.007)	human action
(0.007)	human activity
(0.007)	unit
(0.007)	being
(0.007)	living thing
(0.007)	animate thing
(0.007)	organism
(0.007)	causal agent
	⋮

	cause
(0.013)	abstraction
(0.012)	entity
(0.012)	object
(0.012)	physical object
(0.011)	state
(0.010)	act
(0.010)	human action
(0.010)	human activity
(0.009)	psychological feature
(0.008)	change
(0.008)	cognition
(0.008)	knowledge
(0.008)	noesis
(0.008)	attribute
(0.008)	instrument
(0.007)	event
(0.007)	unit
(0.007)	whole
(0.007)	whole thing
(0.006)	artifact
(0.006)	artefact
(0.006)	activity
(0.006)	relation
(0.006)	status
(0.005)	happening
(0.005)	occurrence
(0.005)	natural event
(0.005)	action
(0.005)	condition
(0.005)	communication
	⋮

	put
(0.015)	entity
(0.015)	location
(0.013)	object
(0.013)	physical object
(0.012)	destination
(0.011)	unit
(0.010)	act
(0.010)	human action
(0.010)	human activity
(0.010)	abstraction
(0.009)	cause
(0.008)	psychological feature
(0.008)	whole
(0.008)	whole thing
(0.008)	artifact
(0.008)	artefact
(0.008)	goal
(0.008)	cognition
(0.008)	knowledge
(0.008)	noesis
(0.008)	change
(0.007)	grouping
(0.007)	group
(0.007)	attribute
(0.007)	being
(0.007)	living thing
(0.007)	animate thing
(0.007)	organism
(0.007)	causal agent
(0.007)	causal agency
	⋮

Fig. 3 Semantic profiles of verbs *pay*, *cause* and *put* for the direct object position

To compare the semantic profiles of two verbs for the same argument position, we measure the divergence between the two probability distributions represented by these semantic profiles. We use a standard divergence measure, Relative Entropy (or KL-divergence), for this purpose:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{12}$$

This measure shows how different the two semantic profiles are, with a value of zero indicating two identical profiles. The divergence of the highly selective verb *pay*

A novel verb	
(0.021)	entity
(0.017)	object
(0.012)	physical object
(0.011)	abstraction
(0.011)	act
(0.011)	human action
(0.011)	human activity
(0.010)	being
(0.010)	unit
(0.009)	living thing
(0.009)	animate thing
(0.009)	organism
(0.009)	cause
(0.009)	causal agent
(0.009)	causal agency
(0.009)	person
(0.009)	individual
(0.009)	someone
(0.009)	somebody
(0.009)	mortal
	⋮

Fig. 4 Semantic profile of a novel verb for the direct object position

from the base profile of Fig. 4 is estimated as 1.4×10^{-3}, whereas the divergence of the less selective verb *put* from the base profile is 4.9×10^{-4}.

3.3 Evolution of Verb Semantic Profiles

Given enough training data, our model can learn appropriate semantic profiles for different verbs. However, because the model learns these profiles from instances of verb usage, we expect each verb profile to go through a gradual generalization process, where it initially reflects the properties of specific verb arguments, and becomes more general over time. For example, upon hearing a couple of usages of a verb such as *I watched a film* and *they watched a movie*, a language learner might assume that the verb *watch* can only be used in the movie-watching context. Similarly, hearing *he ate an orange* and *she is eating an apple* might mislead the learner to think that *eat* can only accept fruits as its direct object. Later and more varied usages of such verbs leads to the formation of a more general profile for their arguments.

We tracked this generalization process for the acquired semantic profiles. As an example, Fig. 5 shows the semantic profile for the direct object position of *make*

After 500 items	
(0.024)	entity
(0.022)	object
(0.022)	physical object
(0.017)	unit
(0.017)	whole
(0.016)	whole thing
(0.016)	artifact
(0.016)	artefact
(0.014)	instrumentality
(0.014)	instrumentation
(0.011)	abstraction
(0.010)	change
(0.010)	act
(0.010)	human action
(0.010)	human activity
(0.010)	device
(0.009)	move
(0.008)	electrical device
(0.008)	flow
(0.008)	course
	⋮

After 5000 items	
(0.015)	act
(0.015)	human action
(0.014)	human activity
(0.014)	abstraction
(0.014)	entity
(0.012)	relation
(0.012)	communication
(0.012)	social relation
(0.011)	object
(0.011)	physical object
(0.010)	event
(0.009)	content
(0.009)	unit
(0.009)	activity
(0.009)	substance
(0.008)	whole
(0.008)	whole thing
(0.008)	artifact
(0.008)	artefact
(0.008)	change
	⋮

Fig. 5 The evolution of the semantic profile for the direct object of *make*

after processing 500 and 5,000 input items. It can be seen that the earlier profile reflects very specific properties (e.g., *electrical device, flow*). The high probability associated with such properties is due to an early use of lexical items such as *filter* and *control* as the direct object of *make*. However, as the model receives more input, the predicted profile reflects more general properties and is even generalized to non-literal usages of *make* (e.g., *make appointment* and *make contribution*, reflected by properties such as *human action* and *communication*).

To monitor the evolution of a semantic profile for each verb, we can compare the semantic profile for an argument position at a given point in learning, and the profile for that position that the model eventually converges to at the end of each simulation. We measure the divergence between the two probability distributions represented by these semantic profiles using the Relative Entropy measure of Eq. (12) (Sect. 3.2). In order to visualize the strength of the selectional preferences for each verb, at each point in time we also compare the divergence of its profile with that of a novel verb (predicted at the same time point). A wide divergence will indicate a stronger preference imposed by the target verb.

Figure 6 shows the profile divergences for the direct object positions of *pay*, *cause* and *put* after processing every 100 input items over a total of 5,000 items. The divergence between the current and end profiles of the same verb is shown by solid lines, and the divergence between the current profiles of the target and the novel verb

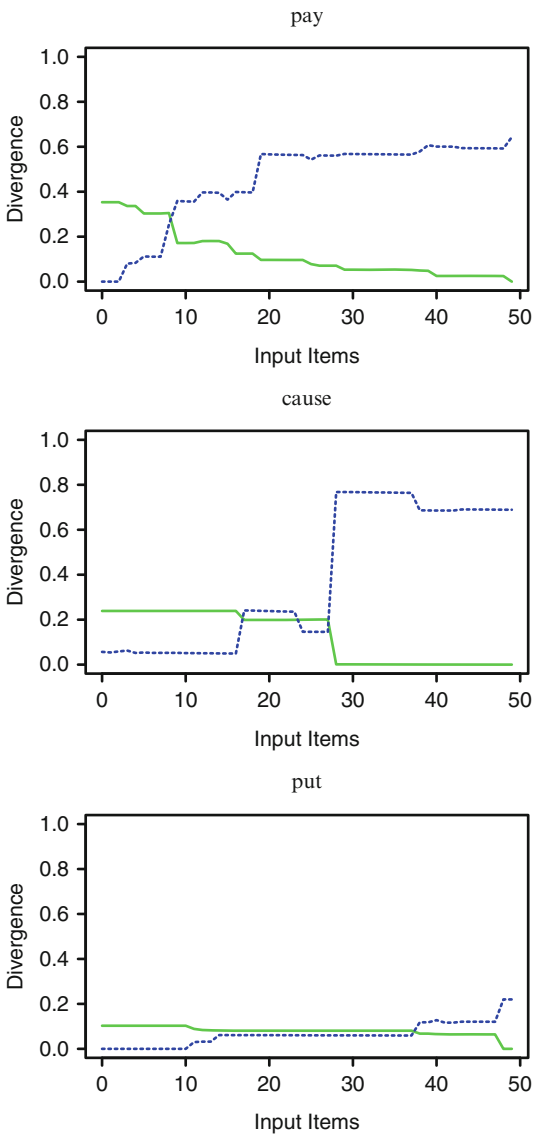

Fig. 6 Learning curves for verb semantic profiles. The x-axis is time (#inputs), and the y-axis is divergence from the profile that the model eventually converges to. *Solid lines* show divergence from the ultimate profile for the same verb, whereas *dashed lines* show divergence from the profile predicted for a novel verb

is shown by dashed lines. Figure 6 shows that more restrictive verbs such as *pay* and *cause* strongly diverge from the profile for a novel verb, and this gap becomes wider as the model processes more input. Since *pay* is much more frequent in our data set than *cause*, its learning curve is smoother and changes more gradually. The profile predicted for *put* (which allows for a wide range of direct objects) remains relatively constant over time, resembling the profile of a novel verb.

Verb	Plausible		Implausible	
see	friend	−32.73	method	−35.94
read	article	−24.24	fashion	−26.68
find	label	−24.74	fever	−26.52
hear	story	−25.75	issue	−25.76
write	letter	−24.47	market	−25.34
urge	daughter	−34.36	contrast	−36.13
warn	driver	−45.99	engine	−45.20
judge	contest	−46.48	climate	−49.43
teach	language	−46.44	distance	−47.47
show	sample	−23.93	travel	−24.40
expect	visit	−46.24	mouth	−45.34
answer	request	−34.49	tragedy	−36.75
recognise	author	−37.85	pocket	−37.88
repeat	comment	−47.08	journal	−48.32
understand	concept	−37.78	session	−37.63
remember	reply	−28.79	smoke	−28.63

Fig. 7 Compatibility scores for plausible vs. implausible verb-noun pairs. Verbs for which the model correctly chooses the plausible argument are shown in boldface

3.4 Verb-Argument Plausibility Judgments

Holmes et al. [14] evaluate verb argument plausibility by asking human subjects to rate sentences like *The mechanic warned the driver* and *The mechanic warned the engine*. Resnik [29] used these data to assess the performance of his model by comparing its judgments of selectional fit against the plausibility ratings elicited from human subjects. He showed that his selectional association measure for a verb and its direct object can be used to select the more plausible verb-noun pair among the two (e.g., <warn,driver> vs. <warn,engine> in the previous example). That is, a higher selectional association between the verb and one of the nouns compared to the other noun indicates that the former is the more plausible pair. Resnik [29] used the Brown corpus as training data, and showed that his model arrives at the correct ordering of more and less plausible arguments in 11 of the 16 cases.

We repeated this experiment using the same 16 pairs of verb-noun combinations. As before, we trained our model on 30,000 extracted frames from the Wall Street Journal corpus (out of 142,000); that is, only one-fifth of the data used by Resnik. For each pair of $<v, n_1>$ and $<v, n_2>$, we calculate the compatibility measure using Eq. (10); these values are shown in Fig. 7. (Note that because these are log-probabilities and therefore negative numbers, a lower absolute value of compatibility(v, n) shows a better compatibility between the verb v and the argument n.) For example, <see,friend> has a higher compatibility score (−32.73) than <see,method> (−35.94). Our model detects 12 plausible pairs out of 16, which is slightly more accurate than Resnik's model. However, these results are reached with a much smaller training corpus.

4 Discussion and Future Directions

In the context of human language learning, it is important to show that the selectional preferences of verbs can be acquired gradually and through online processing of verb usage data. The model presented in this paper for learning and use of verb selectional preferences shows that a semantic profile, or a probability distribution over semantic properties, can be learned for argument positions of individual verbs. Unlike other existing models of selectional preferences, our model is incremental and does not rely on having access to a full hierarchical representation of semantic classes prior to the acquisition of verb selectional preferences. These two properties make our model more cognitively plausible than other existing models. However, there are several directions for improving the model and investigating new applications for it in the future. We will review a few of these directions in the following.

Ambiguity and noise. A central problem for induction of the selectional preferences is noise in the training data. Noise can be due to errors in frame extraction, or due to metaphorical usage of verb-noun compounds. Another problem is word sense ambiguity in the training data which can lead to incorrect generalization, especially because we merge the hypernyms of multiple senses of a word to construct its set of semantic properties. (However, McCarthy and Carroll [22] show that integrating a word sense disambiguation module into a selectional preference induction model does not significantly improve the performance.) Both these problems are expected to be improved by increasing the size of training data and using more sophisticated data processing techniques. Our model shows promise in learning intuitive semantic profiles for verbs and in simulating human plausibility judgements using a small input data set, but processing larger sets of input data can improve the performance of the model. We plan to investigate this in future.

Noun selectional preferences. It has been argued that selectional preferences can be applied for nouns too, since they "prefer" to select certain predicates over others [27]. For example, *cake* prefers to be baked, but not written. We can model this reverse preference as predicting the semantic primitives of the head word (verb) based on the properties of the arguments and the syntactic pattern of the usage. Moreover, the compatibility measure introduced in Eq. (10), Sect. 2.4, can also be used to compare two pairings of a noun with two different verbs, and to measure the preference of that noun for each of the verbs (e.g., ⟨*bake, cake*⟩ vs. ⟨*write, cake*⟩). This possibility of extending to other parts of speech is lacking in the existing models of selectional preferences.

Compound verbs. To our knowledge, computational methods of inducing selectional preferences have only been applied to simple (i.e. single-word) verbs. However, compound verbs such as *take a walk* form a significant proportion of the lexicon in most languages (see [25] in this volume for a computational study on the acquisition of compound verbs). An interesting line of research would be to generalize the current model to learning selectional preferences for compound verbs.

Moreover, in line with the previous proposal, the generalized model can also be applied to compound nouns and the induction of their selectional preferences (similarly, see [10] in this volume for a model of interpreting novel noun-noun compounds).

Wider range of evaluation tasks. Our main goal in this paper was to show through qualitative analysis that it is possible to learn intuitive representations of verb selectional preferences through an incremental process, and to look at the gradual evolution of these representations over time. However, a natural continuation of this work would be to use these acquired preferences in various tasks, similar to what humans do in language learning and processing. The evaluation tasks that have been widely used in the computational models of selectional preferences are mainly of an artificial nature, for example the widely used psudo-disambiguation task [12,31]. We plan to identify more natural tasks for which experimental data from human subjects is available, and compare the performance of our model in these tasks to that of humans.

References

1. Abney, S., & Light, M. (1999). Hiding a semantic hierarchy in a Markov model. In *Proceedings of the ACL workshop on unsupervised learning in natural language processing*. Maryland, USA.
2. Alishahi, A., & Stevenson, S. (2007). A cognitive model for the representation and acquisition of verb selectional preferences. In *Proceedings of the ACL-2007 workshop on cognitive aspects of computational language acquisition* (pp. 41–48). Prague, Czech Republic.
3. Alishahi, A., & Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science: A Multidisciplinary Journal, 32*(5), 789–834.
4. Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes, 25*(1), 50–93.
5. Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.
6. Ciaramita, M., & Johnson, M. (2000). Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proceedings of the 18th international conference on computational linguistics (COLING 2000)*. Saarbrücken, Germany.
7. Clark, S., & Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics, 28*(2), 187–206.
8. Connor, M., Fisher, C., & Roth, D. (2012). Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition*. Springer.
9. Copestake, A., & Briscoe, T. (1991). Lexical operations in a unification-based framework. *Lecture Notes in Computer Science, 627*, 101–119.
10. Devereux, B. J., Costello, F. J. (2012). Learning to interpret novel noun-noun compounds: Evidence from category learning experiments. In *Cognitive aspects of computational language acquisition*. Springer.
11. Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 216–223, Prague, Czech Republic.
12. Gale, W. A., Church, K. W., & Yarowsky, D. (1992). Work on statistical methods for word sense disambiguation. In *AAAI fall symposium on probabilistic approaches to natural language*. Massachusetts, USA

13. Gleitman, L., & Gillette, J. (1995). The role of syntax in verb learning. In P. Fletcher, & B. MacWhinney (Eds.), *Handbook of child language*. Oxford: Blackwell.
14. Holmes, V. M., Stowe, L., & Cupples, L. (1989). Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language, 28*, 668–689.
15. Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT.
16. Johansson, R., & Nugues, P. (2007). Extended constituent-to-dependency conversion for English. In *Proceedings of NODALIDA 2007*, Tartu, Estonia (pp. 105–112).
17. Katz, J., & Fodor, J. (1964). *The structure of language: Readings in the philosophy of language*. Englewood Cliffs, N.J., Prentice Hall.
18. Li, H., & Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics, 24*(2), 217–244.
19. Light, M., & Greiff, W. (2002). Statistical models for the induction and use of selectional preferences. *Cognitive Science, 26*(3), 269–281.
20. MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
21. Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics, 19*(2), 330.
22. McCarthy, D., & Carroll, J. (2003). Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Computational Linguistics, 29*(4), 639–654.
23. Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 17*(3), 235–244
24. Nation, K., Marshall, C. M., & Altmann, G. T. M. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *Journal of Experimental Child Psychology, 86*, 314–329.
25. Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). Child acquisition of multiword verbs: A computational investigation. In *Cognitive Aspects of Computational Language Acquisition*. Springer.
26. Pinker, S. (1994). How could a child use verb syntax to learn verb semantics? *Lingua, 92*, 377–410.
27. Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT.
28. Resnik, P. (1993). *Selection and information: A class-based approach to lexical relationships*. PhD thesis, University of Pennsylvania.
29. Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition, 61*, 127–199.
30. Resnik, P. (1997). Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX workshop on tagging text with lexical semantics: Why, What, and How?* Washington, D.C., USA.
31. Schütze, H. (1992). Context space. In *AAAI fall symposium on probabilistic approaches to natural language*. Massachusetts, USA.
32. Zapirain, B., Agirre, E., & Màrquez, L. (2009). Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, (pp. 73–76). Association for Computational Linguistics.

Index

Acquisition, 31, 39, 45, 46
Acquisition of word categories, 18
Adjective acquisition, 19
Adjectives, 106–108, 110, 121, 123, 125
Agreement, 147
  auxiliary/main verb system, 147
  parsing and tense, 146
Algorithm, 36–38, 46, 48
  clustering algorithm, 87
  incremental algorithm, 87
  unsupervised algorithm, 87
Algorithmic, 46
Algorithmic level, 9
Ambiguity of scenes, 258, 282
Analogy processing, 20
Analysis, 31, 34
Animacy, 222, 225
Argument identification, 285
Argument structure frames, 302
Artificial grammar learning, 139
Artificial intelligence, 7
Artificial neural networks, 11, 106
Auditory primary cortex, 115
Autistic language savants, 160

Basic verbs, 238, 244
Bayesian clustering, 303, 304
Bayesian models, 12
Behavioural methods, 5
Beta distribution, 66
Bipartite network, 56
British National Corpus (BNC), 15
Brown corpus, 15

Canonical predicate-argument, 154
Categories
  abstract categories, 82
  syntactic categories, 81, 82, 84
  word categories, 83, 84
Category learning, 205, 206
CDS. *See* Child-directed speech (CDS)
Child-directed speech (CDS), 83, 84, 94, 268
CHILDES, 14, 17, 29–31, 89, 268
Child language acquisition, 8, 236, 239, 244
Children, 174
Children's productivity, 175
Child's basic patterns, 175
CLAN, 17
Clustering algorithm, 86, 91
Cognitive abilities, 174
Cognitive models, 10, 22
Cognitive plausibility, 2, 9
Cognitive processes, 11
Color, 109, 125
  adjectives, 123
  constancy, 120
  names, 106
Combinatory Categorial Grammar (CCG), 134
Complex networks, 2
Computational algorithms, 16
Computational level, 9
Computational model, 2, 7, 8, 84, 89
Computational models of language, 16
Computational models of language acquisition, 16
Concatenation, 176
Concept relations, 3
Concepts, 199–204, 213, 215–217, 223, 225, 231, 232
Connectionist models, 11
Connectivity, 124, 125
Construction grammar, 174

T. Poibeau et al. (eds.), *Cognitive Aspects of Computational Language Acquisition*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-31863-4, © Springer-Verlag Berlin Heidelberg 2013

Constructions, 302
Co-occurrences, 109
Core knowledge, 107
Corpora, 133
  GENIA corpus, 152
  QuestionBank, 161
Cortical maps, 19
CSLU Toolkit, 45
Cues, 18
  distributional cues, 85
  morphological cues, 85
  phonological cues, 85

Data, 32, 34, 36
Database, 31, 46
Data sparsity, 135
Degree distribution, 58
Dense corpus, 190
Development, 31, 37, 48
Developmental compatibility, 8
Directed speech, 18
Distributional information, 18
Distributional learning, 264
D-structure, 154, 169

Ease of articulation, 53–54
Ease of learnability, 54
Edges, 52
Electroencephalography (EEG), 6
Embryonic syntax, 110
Empiricism, 5
Entities, 109
Errors, 182
Evaluation, 140
Event-related potentials (ERP), 6
Exemplar-based model, 21
Exemplar theory, 202, 204, 213, 216, 223, 224
  GCM, 213–215, 224
Expectation maximization (EM), 271
  variational Bayes EM, 272
Extracted from the input, 175
Eye gaze, 107
Eye-tracking, 6

Fast-mapping, 106, 108
Features
  branching/conjunct parallelism, 149
  head of phrases, 150
  speech categories, 151
Fixed-form, 174, 175
Fixed form patterns, 20
Functional magnetic resonance imaging (fMRI), 6

Gangliar cells, 115
Generalize, 174
Generate language, 189

Hebbian rule, 113
Hidden Markov Model (HMM), 270, 271, 273, 277, 285
HMM. *See* Hidden Markov Model (HMM)
Human language acquisition, 2

Implementation level, 9
Incrementality, 10
Information theory, 13
Innate knowledge, 3
Innateness hypothesis, 4
Intentionality, 107
Interpretation of noun-noun compounds, 21
Invariance, 120
Inversion, 185

Knowledge of language, 135, 167

Language acquisition, 16, 47, 52, 135, 174
Language change, 52
Language innateness, 3
Language modularity, 3
Language variation, 31
Latent BabySRL, 257, 263, 270, 280, 288, 290
Latent training, 274
Laterally interconnected synergetically self-organizing map (LISSOM), 112
Lateral prefrontal cortex, 117
Learnability constraints, 2
Learning form-meaning associations, 20
Learning lexical categories, 18
Learning word meaning, 18
Lexical development norms, 286
Light verb, 238
Light verb constructions, 236
Linguistic competence, 136
Linguistic constructions
  active, 155
  auxiliary/main verb system, 153
  declarative, 153
  interrogative, 153
  parsing passives, 154
  small clause, 157
  subject-auxiliary verb questions, 162
  wh-movement, 153
Linguistic knowledge, 17

Linguistic regularization, 154
Logical form, 154, 159
Long-distance displacements, 160

Machine learning, 16
Machine learning techniques, 10
Macroscopic, 52
Manchester corpus, 177
manner of articulation, 55
Mapping of words to meanings, 19
Markedness hierarchy, 55
Marr, 9
Maximal perceptual contrast, 53
MDL. *See* Minimum description length (MDL)
Meanings, 18
Medial geniculate nucleus, 115
Memory, 20
Mesoscopic, 52
Microscopic, 51
Minimum description length (MDL), 13
Model, 19, 85
  incremental model, 85
Model-testing, 48
Morphology, 17, 18
MRC, 90
Multiword expressions, 21
Multiword lexeme, 236–238, 243, 247–252
Multi-word utterances, 176
Multiword verbs, 237

Nativism, 3, 5
Nodes, 52
Non-literal expressions, 239, 240, 242, 244–247, 250–252
Non-local dependencies, 20
Noun acquisition, 19
Noun-noun compounds, 199, 202, 204, 213, 216, 220, 221, 223, 225, 230–232
  CARIN model, 200, 203, 215, 217
  concept-based approaches, 201, 231
  relation-based approaches, 200, 201, 203, 231
Nouns, 18, 106, 108, 109, 121, 123, 125
Novel word, 93
Nurture, 5

Output, 182

Parameter estimation, 160
Parsers, 133
  Stanford, 150
Passive errors, 135
Passives, 20, 138
PCFG. *See* Probabilistic context free grammars (PCFG)
Penn Tree (PTB), 133
Perception, 20
Perceptually driven associational learning, 19
Phon, 17, 29, 44
PhonBank, 17, 29, 31
Phonetic, 31, 34, 35, 45
Phonological, 18, 30, 47
Phonological rehearsal loop, 117
Phonology, 30, 47
Plausibility judgments, 313
Population code, 122
Population coding, 120, 121
PoStags, 86
Poverty of the stimulus, 4
Power-law, 60
PP attachments, 156
Pragmatics, 3
Predicate-argument structure, 20
Predicates, 109
Predict, 305
Preferential attachment, 60
Prefrontal cortex, 110
Primary auditory cortex, 116
Primary linguistic data, 4
Probabilistic context free grammars (PCFG), 12
Probabilistic framework, 13
Probabilistic modeling, 12
Problem-solving, 203
PropBank, 267, 268
Properties of words, 18
PTB, 167
ptIn, 240

QuestionBank, 141, 165
Question inversion, 19

Ray Jackendoff, 146
  text reading machine problem, 147
Reading times, 6
Receptive fields, 115
Regularizations, 20
Relations, 109

Saliency, 108
Selectional preferences, 22
Self-organization, 106
Self-organizing, 54

Self-organizing maps, 111
Semantic role labeling (SRL), 261, 262, 267, 270, 277
Semantic roles, 21
  classification, 21
Semantics, 17
  bootstrapping, 259, 290
  interpretations, 20
  prediction, 98
  profile, 22, 301, 307, 310
  properties, 22
  relations, 200–203, 205, 213–216, 223, 225, 232
Shape, 124
Shape bias, 107
Similarity, 87
Simple generative mechanisms, 191
Simple recurrent network (SRN), 265, 266
Single-word utterances, 176, 190
Slot-and-frame, 184
Social conventions, 3
Social cue, 14
Software, 30
Spontaneous retinal activity, 119
SRL. *See* Semantic role labeling (SRL)
SRN. *See* Simple recurrent network (SRN)
Statistical cues, 4
Statistical parsers
  attachment errors, 154
  markovization, 152
  over-lexicalization, 151
  parameter estimation, 153
  small sampling effects, 151
  successive state-splitting, 152
Statistical parsing, 134
Structured prediction, 274
Structure-mapping, 260, 261, 284
Subcategorization frames, 203
Superior temporal sulcus, 115
Switchboard corpus, 15
Symbolic modeling, 10
Syntactic bootstrapping, 260, 261, 284, 290
Syntactic burst, 20, 173
Syntactic categories, 18
Syntactic fixedness, 242, 243
Syntactic forms, 20
Syntax, 17, 18, 204

TalkBank, 14
Target word, 175
Tense
  imperative, 148
  marking, 20
Terms, 109
Testing, 45, 140
Testing procedure, 176
T-marking, 138
Tomasello, 190
Tonotopic, 120
Training data, 140
Training data augmentation, 143
  QuestionBank, 143
Transcript, 177, 185
Transcription, 32, 34, 36
Tree-level transformations, 154

UCLA Phonoogical Segment Inventory Database (UPSID), 54
Universal grammar (UG), 4
Unnatural language constructions, 139
  mirror reversed, 160
Unsupervised Part of Speech (POS), 271, 273
UPSID. *See* UCLA Phonoogical Segment Inventory Database (UPSID)
Usage-based, 20, 84, 174
Usage-based theories of language acquisition, 5
U-shaped generalization curve, 8

Ventral stream, 113
Verb-argument compatibility, 306
Verbs, 18, 22, 108, 203
Verb selectional preferences
  computational models, 299
  computational study, 298
Visual pathway, 113
Visual World Paradigm, 3
Vocabulary growth, 8

Wexler and Culicover, 169
Whole object bias, 107
Wh-questions, 20, 138, 141
Word learning, 8, 236, 237, 243, 247, 248, 250, 252
WordNet, 91, 226, 265
Word order, 187
Word order errors, 183
Working memory, 110, 117, 118, 123, 125
WSJ, 164

Young children, 182

Zipf-distributional, 134
Zipf's-law, 60