

Studies in
NATURAL
LANGUAGE
PROCESSING

LANGUAGE, COGNITION, AND COMPUTATIONAL MODELS



Edited by
Thierry Poibeau
Aline Villavicencio



CAMBRIDGE

Language, Cognition, and Computational Models

How do infants learn a language? Why and how do languages evolve? How do we understand a sentence? This book explores these questions using recent computational models that shed new light on issues related to language and cognition. The chapters in this collection propose original analyses of specific problems and develop computational models that have been tested and evaluated on real data. Featuring contributions from a diverse group of experts, this interdisciplinary book bridges the gap between natural language processing and cognitive sciences. It is divided into three sections, focusing respectively on models of neural and cognitive processing, data-driven methods, and social issues in language evolution. This book will be useful to any researcher and advanced student interested in the analysis of the links between the brain and the language faculty.

THIERRY POIBEAU is Director of Research at CNRS and head of the LaTTiCE laboratory in Paris, France. He is also an affiliated lecturer at the Department of Theoretical and Applied Linguistics (DTAL) of the University of Cambridge. He works on natural language processing (NLP), in particular on information extraction, question answering, semantic zoning, knowledge acquisition from text, and named entity tagging.

ALINE VILLAVICENCIO is affiliated with the Institute of Informatics, Federal University of Rio Grande do Sul in Brazil, and with the School of Computer Science and Electronic Engineering, Essex University, UK. She is a fellow of CNPq (Brazil). Her research interests in natural language processing are in computational models of acquisition of linguistic information from data, distributional semantic models, multiword expression, and applications like text simplification and question answering.

Studies in Natural Language Processing

Volumes in the SNLP series provide comprehensive surveys of current research topics and applications in the field of natural language processing (NLP) that shed light on language technology, language cognition, language and society, and linguistics. The increased availability of language corpora and digital media, as well as advances in computer technology and data sciences, has led to important new findings in the field. Widespread applications include voice-activated interfaces, translation, search engine optimization, and affective computing. NLP also has applications in areas such as knowledge engineering, language learning, digital humanities, corpus linguistics, and textual analysis. These volumes will be of interest to researchers and graduate students working in NLP and other fields related to the processing of language and knowledge.

Chief Editor:

Chu-Ren Huang – The Hong Kong Polytechnic University,
Department of Chinese and Bilingual Studies

Associate Editor:

Qi Su – Peking University, School of Foreign Languages

Editorial Board:

Steven Bird – University of Melbourne, Department of Computing
and Information Systems

Francis Bond – Nanyang Technological University,
Division of Linguistic and Multilingual Studies

Alessandro Lenci – Università di Pisa, Dipart. di Filologia,
Letteratura e Linguistica

Lori Levin – Carnegie Mellon University, Language
Technologies Institute

Maarten de Rijke – University of Amsterdam, Informatics Institute

Nianwen Xue – Brandeis University, Computer Science Department

Language, Cognition, and Computational Models

Edited by

Thierry Poibeau

CNRS, Paris, France

Aline Villavicencio

Universidade Federal do Rio Grande do Sul, Brazil and University of Essex, UK



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE

UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi - 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107162228

DOI: 10.1017/9781316676974

© Cambridge University Press 2017

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2017

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-16222-8 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party Internet Web sites referred to in this publication
and does not guarantee that any content on such Web sites is, or will remain,
accurate or appropriate.

Contents

<i>Figures</i>	<i>page</i>	ix
<i>Tables</i>		xi
<i>Contributors</i>		xv

Part I About This Book

1	Introduction: Cognitive Issues in Natural Language Processing	3
	THIERRY POIBEAU AND ALINE VILLAVICENCIO	
1.1	On the Relationships between Natural Language Processing and Cognitive Sciences	3
1.2	Recent Issues in Cognitive Aspects of Language Modeling	8
1.3	Content and Structure of the Book	14

Part II Models of Neural and Cognitive Processing

2	Light and Deep Parsing: A Cognitive Model of Sentence Processing	27
	PHILIPPE BLACHE	
2.1	Introduction	27
2.2	An Interdisciplinary View of Language Processing	28
2.3	The Theoretical Framework: Property Grammars	39
2.4	Chunks, Constructions, and Properties	42
2.5	The Hybrid Architecture	46
2.6	Conclusion	49
3	Decoding Language from the Brain	53
	BRIAN MURPHY, LEILA WEHBE, AND ALONA FYSHE	
3.1	Introduction	53
3.2	Grounding Language Architecture in the Brain	55
3.3	Decoding Words in the Brain	62
3.4	Phrases in the Brain	66
3.5	Stories in the Brain	68
3.6	Summary	72

4	Graph Theory Applied to Speech: Insights on Cognitive Deficit Diagnosis and Dream Research NATÁLIA BEZERRA MOTA, MAURO COPELLI, AND SIDARTA RIBEIRO	81
4.1	Introduction	82
4.2	Semantic Analysis for the Diagnosis of Psychosis	84
4.3	What Is a Speech Graph?	85
4.4	Speech Graphs as a Strategy to Quantify Symptoms on Psychosis	88
4.5	Differences in Speech Graphs due to Content (waking × dream reports)	92
4.6	Speech Graphs Applied to Dementia	94
4.7	Future Perspectives	96
Part III Data Driven Models		
5	Putting Linguistics Back into Computational Linguistics MARTIN KAY	101
5.1	Explicit and Implicit Information	101
5.2	Features	108
5.3	Linguistic Computation and Computational Linguistics	114
5.4	Conclusion	116
6	A Distributional Model of Verb-Specific Semantic Roles Inferences GIANLUCA E. LEBANI AND ALESSANDRO LENCI	118
6.1	Representing and Acquiring Thematic Roles	119
6.2	Characterizing the Semantic Content of Verb Proto-roles	122
6.3	A Distributional Model of Thematic Roles	130
6.4	Experiments with Our Neo-Davidsonian Model	139
6.5	Conclusion	148
7	Native Language Identification on EFCAMDAT XIAO JIANG, YAN HUANG, YUFAN GUO, JEROEN GEERTZEN, THEODORA ALEXOPOULOU, LIN SUN, AND ANNA KORHONEN	159
7.1	Introduction	159
7.2	Data	165
7.3	Methods	168
7.4	Results	172
7.5	Conclusion	181
8	Evaluating Language Acquisition Models: A Utility-Based Look at Bayesian Segmentation LISA PEARL AND LAWRENCE PHILLIPS	185
8.1	Introduction	185
8.2	Early Speech Segmentation	187

Contents	vii
----------	-----

8.3 A Bayesian Segmentation Strategy	190
8.4 How Well Does This Work Cross-Linguistically?	196
8.5 How Useful Are the Units?	207
8.6 Closing Thoughts	219

Part IV Social and Language Evolution

9 Social Evolution of Public Languages: Between Rousseau's <i>Eden</i> and Hobbes' <i>Leviathan</i>	227
ANNE REBOUL	
9.1 Introduction	227
9.2 Is Language a Communication System in the Strong Sense?	228
9.3 What is the Proper Social Account for the Exaptation of Language for Communication?	233
9.4 Conclusion	250
10 Genetic Biases in Language: Computer Models and Experimental Approaches	256
RICK JANSSEN AND DAN DEDIU	
10.1 Introduction	256
10.2 Computer Models of Cultural Evolution	262
10.3 Cultural Feedback	278
10.4 Conclusion	281
11 Transparency versus Processing Efficiency: A Case Study on German Declension	289
REMI VAN TRIJP	
11.1 Introduction	289
11.2 German Declension: Not as Awful as It Seems	290
11.3 Evaluating the Efficiency of Syncretism	300
11.4 Discussion and Conclusions	314
Index	319

Figures

2.1 A classical generative architecture of language processing	page 29
2.2 Output of the Stanford parser	30
2.3 The main ERP components in language processing	31
2.4 General organization of the three-phases model [Friederici, 2011]	32
2.5 The differences between conditions in idiomatic context (COR, correct sentence; REL, soft violation; UNREL, hard violation)	35
2.6 Parafoveal vision. Extracting features from the surrounding words	36
2.7 Early EEG effects of syntactic violation (mismatch negativity)	37
3.1 An overview of the location and timing of language processing (for sentences) in the brain	56
3.2 An example MEG recording averaged over twenty repetitions of a person reading the word <i>bear</i>	58
3.3 An fMRI image averaged over six repetition of a person reading the word <i>bear</i>	59
3.4 Several slices from fMRI images showing the <i>learned</i> proportion of brain activation that can be associated with a particular verb from the set of twenty-five verbs used in Mitchell et al. (2008)	64
3.5 The predicted (<i>top row</i>) and observed (<i>bottom row</i>) brain activation for a particular person while reading the word “celery” or “airplane” (<i>left and right columns, respectively</i>)	65
3.6 Story reading brain map, adapted from Wehbe et al. (2014b)	70
3.7 Time-line of word integration across the MEG helmet	71
4.1 Examples of speech graphs from dream reports of schizophrenic, bipolar, and control subjects	86
4.2 Examples of Speech Graph Attributes	87
4.3 Linear correlation between SGA and word count (WC)	89
4.4 Representative speech graphs extracted from dream reports from a schizophrenic, a bipolar and a control subject	91

4.5 Representative speech graphs examples extracted from dream and waking reports from the same schizophrenic, bipolar, and control subject	93
6.1 The verb role description interface in Crowdflower	125
6.2 Proportion of verb-feature pairs in a ± 2 -sentences window, modulated by application of different frequency thresholds	143
6.3 Precision of the different models evaluated against the dataset of extended role-based features	145
6.4 Precision of the different models evaluated against the dataset of role-based features	146
7.1 Example of unlexicalized and lexicalized version of production rule features (PR) of sentence “I speak English”	170
7.2 Example of unlexicalized and lexicalized version of dependency features (Depd) on sentence “I come from western Europe”	171
7.3 Performance of different configuration variations of word 1-gram feature at levels 1–3	173
7.4 Performance of different variations of character 5-gram at levels 1–3	174
7.5 Performance of different POS n-grams at levels 1–3	174
7.6 Performance of unlexicalized and lexicalized syntactic relational features at levels 1–3	175
8.1 Plate diagram of the Intentional strategy’s generative model	214
10.1 The IL social structure	263
10.2 A generalization of the phases observable in a typical IL run	264
11.1 Parsing of the utterance <i>Die Kinder gaben der Lehrerin die Zeichnung</i> “The children gave the drawing to the (female) teacher”	294
11.2 A type hierarchy proposed for German case agreement	296
11.3 The search tree for <i>Die Kinder gaben der Lehrerin die Zeichnung</i> using feature matrices in the grammar	300
11.4 This chart compares Old High German (black) and New High German (white) in how many utterances were disambiguated during parsing in four different analyses	305
11.5 These spider charts each take an article at their center and then show the interpolated distances between that article and other forms of the same paradigm	314

Tables

4.1 Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2012)	<i>page</i> 90
4.2 Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2014)	92
6.1 Distinct Slot-Based Features and Consistency for Verb-Role Pair	129
6.2 Distinct features in the gold standard datasets	130
6.3 Top 10 associated features per role extracted with the full model	149
6.4 Evaluation against the dataset of extended role-based features	150
6.5 Evaluation against the dataset of role-based features	150
7.1 The statistics of EFCAMDAT	166
7.2 Top 10 nationalities by the number of writings	166
7.3 Number of writings at 16 proficiency levels	166
7.4 Example topics for 16 proficiency levels	167
7.5 A subset of EFCAMDAT used in this work	168
7.6 Example of three POS n-gram subtypes	169
7.7 Performance of each individual feature type	175
7.8 Accuracy gain (+%) or loss (-%) of leave-one-out experiments	176
7.9 The top 5 features for different feature types and proficiency levels	177
7.10 Example of unique punctuations used in particular class: Brazilians(br) make more mistakes of replacing a quote mark with an acute accent mark	178
7.11 Chinese Learners tend to underuse dash	178
7.12 Example of phrases frequently used by particular class	179
8.1 Summary of the syllabified child-directed speech corpora, including the CHILDES database corpora they are drawn from (Corpora), the age ranges of the children they are directed at (Age range), the number of utterances (# Utt), the number of unique syllables (# Syl types), the average number	

of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob)	196
8.2 Best free parameter values for all unigram and bigram Bayesian segmentation strategies across each language	199
8.3 Word token F-scores for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese	200
8.4 Average NSE scores across all utterances in a language's corpus, ordered from lowest to highest NSE and compared against the idealized inference token F-score for the language	202
8.5 Percentage of errors that resulted in an oversegmentation as compared to adult orthographic segmentation	203
8.6 Examples of reasonable errors (with English glosses) made in different languages	204
8.7 Percentage of model errors that produced reasonable errors of different kinds	205
8.8 Adjusted word token F-scores, counting reasonable errors as acceptable output, for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese	206
8.9 Extrinsic measure evaluations for the joint model from Doyle & Levy (2013) (DP-Bi+Stress) compared against the original DP-Bi model	210
8.10 Downstream evaluation for different DP segmentation strategy variants, compared against the adult orthographic segmentation and the random oracle baseline in English, German, and Hungarian	211
8.11 Extrinsic measure evaluations for the joint model (+Intention, +Intention+Only1) from Johnson et al. (2010) compared against the original DP segmentation strategies in isolation (Base)	216
8.12 Segmentation and word-object mapping results for modeled learners in Phillips & Pearl (2015a)	217
11.1 German definite articles are marked for case, number, and gender	291
11.2 German indefinite articles follow roughly the same declension pattern as the definite articles	292
11.3 Attributive adjectives take three different declension patterns depending on the presence and kind of a preceding determiner	293
11.4 The feature matrix for German case	298
11.5 The feature matrix for <i>der</i>	299
11.6 The feature matrix for <i>Lehrerin</i>	299
11.7 The feature matrix for <i>der Lehrerin</i>	299
11.8 The Old High German definite article paradigm	303

11.9	The discrete representation of the NHG definite articles <i>die</i> and <i>das</i> in sets of distinctive features per phoneme	309
11.10	The articulatory effort for each definite article on the left, and its ease of articulation to the right	311
11.11	Articulatory distinctiveness with respect to its nearest phonological neighbor on the left, and the same measure interpolated on a scale from zero to ten to the right	312

Contributors

THEODORA ALEXOPOULOU Department of Theoretical and Applied Linguistics, University of Cambridge, UK

PHILIPPE BLACHE Laboratoire Parole et Langage, CNRS & Aix-Marseille Université, France

MAURO COPELLI Physics Department, Federal University of Pernambuco, Brazil

DAN DEDIU Language and Genetics Department, Max Planck Institute for Psycholinguistics, The Netherlands

ALONA FYSHE Department of Computer Science, University of Victoria, Canada

JEROEN GEERTZEN Department of Theoretical and Applied Linguistics, University of Cambridge, UK

YUFAN GUO IBM Research, USA

YAN HUANG Department of Theoretical and Applied Linguistics, University of Cambridge, UK

RICK JANSSEN Language and Genetics Department, Max Planck Institute for Psycholinguistics, The Netherlands

XIAO JIANG Computer Laboratory, University of Cambridge, UK

MARTIN KAY Department of Linguistics, Stanford University, USA

ANNA KORHONEN Department of Theoretical and Applied Linguistics, University of Cambridge, UK

GIANLUCA E. LEBANI Department of Philology, Literature, and Linguistics, University of Pisa, Italy

ALESSANDRO LENCI Department of Philology, Literature, and Linguistics, University of Pisa, Italy

NATÁLIA BEZERRA MOTA Brain Institute, Federal University of Rio Grande do Norte, Brazil

BRIAN MURPHY Centre for Data Science and Scalable Computing, Queen's University, Belfast, Northern Ireland

LISA PEARL Departments of Linguistics and Cognitive Sciences, University of California, Irvine, USA

LAWRENCE PHILLIPS Department of Cognitive Sciences, University of California, Irvine, USA

THIERRY POIBEAU Lattice laboratory, CNRS and, Ecole Normale Supérieure and Université Sorbonne nouvelle, France

ANNE REBOUL Institute for Cognitive Sciences-Marc Jeannerod, CNRS UMR 5304, University of Lyon, France

SIDARTA RIBEIRO Brain Institute, Federal University of Rio Grande do Norte, Brazil

LIN SUN Greedy Intelligence, China

REMI VAN TRIJP Sony CSL Paris, France

ALINE VILLAVICENCIO Institute of Informatics, Federal University of Rio Grande do Sul, Brazil, and School of Computer Science and Electronic Engineering, University of Essex, UK

LEILA WEHBE Helen Wills Neuroscience Institute, University of California, Berkeley, USA

Part I

About This Book

1 Introduction: Cognitive Issues in Natural Language Processing

Thierry Poibeau and Aline Villavicencio

1.1 On the Relationships between Natural Language Processing and Cognitive Sciences

This introduction aims at giving an overview of the questions and problems addressed jointly in natural language processing and cognitive science. More precisely, the idea of this introduction, and more generally of this book, is to address how these fields can fertilize each other, bringing recent advances to produce richer studies.

Natural language processing is fundamentally dealing with semantics and more generally with knowledge. Cognitive science is also mostly dealing with knowledge: how knowledge is acquired and processed in the brain. The two domains have developed largely independently, as we discuss later in this Introduction, but there are obvious links between the two, and a large number of researchers have investigated problems involving the two fields, in either the data or the methods used.

1.1.1 A Quick Historical Overview

The landscape of natural language processing (NLP) has dramatically changed in the last decades. Until recently, it was generally assumed that one first needs to adequately formalize an information context (for example information contained in a text) in order to be able to subsequently develop applications dealing with semantics (see, for example, Sowa 1991; Allen 1994; Nirenburg and Raskin 2004). This initial step involved manipulating large knowledge bases of manually hand-crafted rules, and has resulted in the new field of “knowledge engineering” (Brachman and Levesque 2004).

Knowledge can be seen as the result of the confrontation of our a priori ideas with the reality of the outside world. This leads to several difficulties: (1) the task is potentially infinite since people constantly perceive a multiplicity of things; (2) perception interferes with information already registered in the brain, leading to complex inferences with commonsense knowledge; (3) additionally, very little is known about how information is processed in the brain, which makes things even harder to formalize.

To answer some of these issues, a common assumption is that knowledge could be disconnected from perception, which led to projects aiming at developing large static databases of “common sense knowledge” from CYC (Lenat 1995) to more recent general domain ontologies like ConceptNet (Liu and Singh 2004). However, these projects have always led to databases that, despite their sizes, were never enough to completely and accurately formalize a given domain, and domain-independent applications were thus even more unattainable. Moreover, very quickly different problems appeared since contradicting facts, variable points of view and subjective information cannot be directly formalized in a static database aiming to provide a general, multipurpose source of “ground-truth” information.

Despite these issues, a formalization of the textual content has often been the basis of most treatments for more than fifty years, since the beginning of NLP as a field in the late 1940s, with the creation of the first computers, to the late 1990s (Jurafsky and Martin 2000). Things have gradually changed in the last twenty to twenty-five years, for two main reasons: (1) the power of modern computers, capable of providing extensive calculation capacities and storing amazingly large collections of data, and (2) the availability of data through the Web, which provides an unseen and constantly expanding collection of text, images, and videos that goes far beyond anything people imagined before. Current corpora contain several billion words, leading to new discoveries “just” by finding patterns revealed by automatic means. As for the text-processing domain, machine learning approaches (Manning and Schütze 1999) are capable of discovering rare word configurations and rare correlations, leading to constant progress and better performances, even for rare events.

Machine learning approaches are now prominent and achieve the best results on most tasks, including when semantics is at stake. Empirical approaches are generally geared toward practical tasks (e.g., parsing or machine translation) and most of the time do not implement any specific theory, let alone any cognitive considerations.

As a consequence, cognitive science and NLP have both evolved quite independently in the last three decades. On the one hand, NLP has made impressive progress in most tasks. Performance of complex systems can now be considered as satisfactory for some tasks and some specific needs, even if the results are still far from perfect. One example is the IBM Watson system that won the television game show *Jeopardy!*¹ a few years ago (Ferrucci 2012). On the other hand, cognitive science has also made much progress, leading to new insights in our knowledge of language processing in the brain.

¹ Jeopardy is an American TV game similar to a question-answering task, except that candidates have to find questions corresponding to answers, rather than answers corresponding to questions.

So, why should we try to establish any link between these two domains if they have largely evolved apart from each other? Is it even relevant to try to establish links? We think we should answer positively to these questions since new connections can be established between the two fields. NLP now widely uses new methods based on machine learning techniques. Even if these methods are not linked to any specific cognitive theory, they provide the means for extracting relevant information from large masses of data. This could be compared to the activity of the human brain extracting information all the time from the different channels of perception.

Ongoing research in both domains can also be enlightening for the other domain. For example distributional models are relevant for semantic theory, as well as for cognitive theories. These results can be used in studies related to lexical structure and lexical storage in the brain, and in more applied fields such as language acquisition and language pathology studies.

The opposite is also true: cognitive science has largely adopted computational models as a way to formalize, test, and validate existing theories, especially in the field of language comprehension. One of the main goals of language comprehension is to elaborate predictive models of “language complexity” often through what is called “surprisal effect”: a linguistic sequence is more or less complex depending on its structure and on the frequency of lexical items used to form a sentence. The next word of a sequence can be predicted more or less accurately (with more or less “surprise”) and traditional hypotheses can now be tested and validated with computational models. These models complement traditional methods, including real tests and neuroimaging experiments (especially based on electroencephalograms) by providing a sound and formal basis for these previous proposals.

1.1.2 Artificial and Natural Systems

There is a long tradition of reflecting on the relationship between natural and artificial systems (Holland 1992). In artificial intelligence, often the goal is not to directly reproduce how information is processed in the brain, since there is a clear lack of knowledge on how the brain works. Rather, scientists and philosophers are more interested in the results obtained by artificial systems. For example, we can ask ourselves: To what extent can a machine produce valuable results for tasks such as text production or, more specifically, translation or dialogue? Is it possible to converse with a machine without noticing that the interlocutor is a machine and not a human? Is it possible to get a translation produced by a machine and not realize that the translation was not made by a human being? In other words, to what extent is it possible to reproduce on a machine tasks involving some kind of “intelligence”?

These are exactly the kinds of questions Turing was dealing with in the 1940s and which led him to propose the famous Turing test. The test states that if a machine is able to converse with a human without him noticing he is speaking with a machine, then this is the sign that the computer has some form of intelligence (Turing 1950; Moor 2003).

There have been numerous discussions on the validity of the Turing test, the key point being whether a dialogue is enough to prove the existence of intelligence. Although even very simple systems are capable of generating seemingly interesting interactions, the levels of real “understanding” shown by these machines are extremely limited. Let’s take the famous example of ELIZA the dialoging system developed by Weizenbaum in 1966 (Weizenbaum 1966). This system was able to simulate a dialogue between a psychotherapist and a patient. ELIZA was in fact just a series of regular expressions to derive questions from the patient’s utterances. It was able to produce, for example, the question “Why are you afraid of X?” from the sentence “I am afraid of X.” The system also included a series of ready-made sentences that were used when no predefined patterns were applicable (for example, “Could you specify what you have in mind?,” “Tell me more,” or “Really?”). Despite its simplicity, ELIZA enjoyed great success, and some patients really thought they were conversing with a real doctor through a computer.

One important point is that artificial systems do not need to directly reproduce human strategy to perform different tasks involving knowledge and reasoning. They do not even have to perform the task, only give the illusion that it is performed. ELIZA has proved that very simple systems are enough to deceive human beings to some extent.² On the other hand, although humans may listen and answer superficially to their interlocutor in a conversation, in general they do perform these tasks using their brains and keeping track of the information expressed throughout the dialogue, not just producing ready-made sentences without taking most of the previous exchanges into consideration.

Additionally, given the lack of knowledge about processes in the brain, the idea of drawing parallels between artificial and natural systems has never been seriously pursued. On the contrary, it has always been largely criticized. At best, artificial systems implement a specific theory but most often they just use a pragmatic approach, where what matters is the final result, not the way it is obtained (like with ELIZA).

The 1990s popularized the use of machine learning methods for language processing (Mitchell 1997; Manning and Schütze 1999). In this approach most of the knowledge comes from corpora, with the assumption that they are now

² This was at a time when machines were rare and people were more technologically naive than today, less used to interacting with computers. However, with some adaptation, the same kind of experiment could probably be reproduced in the present day.

large enough to provide sufficient evidence for further processing. In this case, the goal consists of identifying regularities in data that are likely to also be found in new data. The initial corpus (the training data) must be representative of the domain and should provide enough evidence so as to allow a good coverage of the data to be analyzed in the future. All NLP tasks have been affected by this new approach, which currently dominates the field. For example, there are enough bilingual corpora available to directly train machine translation systems that use this data to translate new texts. There are also large treebanks with syntactically annotated texts that can be used to automatically construct statistical parsers trained on the data.

In this context, it is interesting to look at the role of semantics in the most recent generation of NLP systems. Generally, semantics and statistics are seen as opposites: on the one hand, content representation; on the other hand, computation. However, this opposition is far too simplistic. For example, statistics allows one to accurately characterize degrees of similarity between words or between documents (Pantel and Lin 2002; Turney and Pantel 2010), capturing some semantics. Statistics also offer powerful ways of representing word senses through a precise analysis of the usage of words in context, thanks to distributional models or more recently to the highly popular word embeddings (Fu et al. 2014). One issue with these models is that there is no clear definition of what semantics or even a word sense or a definition is. When we look at traditional dictionaries, it is immediately apparent that the number of senses per word differs: some dictionaries are more detailed, some more general, depending on their goal and their audience. However, word senses are not always mutually exclusive, and often different senses can be used to explain a word in context. The notion of graded word sense has been proposed to characterize this state of affairs: several word senses could apply to the same occurrence of a word, with different degrees of accuracy. In fact, Erk and McCarthy (2009) proved that different definitions of ambiguous words can be more or less relevant at the same time. For example, “paper” could refer to “report,” “publication,” or “medium for writing,” which are all supposed to be different word senses according to a reference dictionary. Therefore, word senses are not disjunct but largely overlap, and often more than one word sense would be simultaneously activated for a given context. Statistics, by characterizing word meanings from usage and context, gives complex representations that are very different from traditional ones, but can nevertheless be compared with cognitive models and may give more relevant results than previously.

Statistics are also relevant to characterize idioms, frozen expressions, and even equivalencies between languages. In the last case, bilingual corpora can be aligned at the sentence or even at the word level (Tiedemann 2011). It is possible to observe more or less regular equivalencies between words in context and compute similarities between sequences of variable length

(m-n equivalencies, like between “potato” and “pomme de terre” or “kick the bucket” and “passer l’arme à gauche,” to use two examples in English and French). Because computers can automatically process very large bilingual corpora (several millions or even a billion words), it is now possible to get richer and infinitely more precise bilingual dictionaries than before. The structure of these dictionaries should be studied from a cognitive point of view: they are interesting since they are obtained from raw data given as input, without a priori knowledge and with no predefined theory. In this sense this sense that we can say that artificial models based on machine learning methods encode some form of meaning that may make sense from a cognitive point of view.

The last decade has seen an impressive amount of research aiming at linking text and image (Chrupala et al. 2015); the general idea is that the acquisition of word meaning is a multimodal process in which vision plays a major role. The issue is then to provide rich enough representations of multimodal input, since it is already difficult to provide a relevant and cognitively plausible representation of an image or a video. Large online collections of images along with their metadata have been used as way to develop models of acquisition of lexical meaning. One particularly interesting point in these experiments is that meaning is not fixed and is gradually associated with specific objects through the observation of regularities in the environments. However, despite this interest, metadata contributes to these results, providing specific information, and the relationship between this kind of experiment and human perception should be questioned.

1.2 Recent Issues in Cognitive Aspects of Language Modeling

1.2.1 NLP and Language Comprehension

Language comprehension has always been an important research issue in both linguistics and cognitive science. For example, the 1970s and 1980s saw a great amount of research on grammar formalisms (Francez and Wintner 2012). One issue was to show the plausibility of these models in terms of language comprehension (Bergen and Chang 2003). These formalisms also provided information on language complexity, for example by highlighting equivalencies between linguistic phenomena and the formal devices needed to represent them (Joshi 1990).

Recent research has shown the convergence of models coming from both sides, linguistics and cognitive science. For example, the very popular ACT-R theory (Adaptive Control of Thought–Rational) has been developed since the early 1970s by John Anderson (Anderson 1976; Anderson et al. 2004). The theory aims to explain how the brain works in a modular way and how these modules interact to make comprehension possible. Applied to language,

it means that different parts of a sentence are analyzed by different modules and calculating the meaning of the sentence corresponds to assembling the various pieces of information stored by these different modules (Gee and Grosjean 2004; Anderson et al. 2004). Partly independently, in NLP, since the 1990s chunks also play a major role. As defined by Abney (1991), they generally consist of “a single content word surrounded by a constellation of function words, matching a fixed template” (e.g., a noun chunk). These simple units can be described by context-free grammar (CFG), whereas the structure of sentences (or, in other words, the relations between chunks) corresponds to more complex schemas that cannot be described by simple CFGs. Blache (2013) proposes to apply this theory to parsing. Blache proposes an adapted version of the activation function that takes advantage of the representation of linguistic information in terms of low-level features including frequency information. This model subsequently simplifies the analysis of sentences in accordance with cognitive models.

A related line of research also based on the ACT-R theory investigates the notion of comprehension by focusing on comprehension difficulty (Levy 2013). There is a consensus in the comprehension community to explain comprehension with the support of two main notions: memory and expectation. *Memory* refers to the ability to store information and bring it to the front whenever necessary. *Expectation* refers to the predetermination of a lexical category depending on the context. Levy (2013) discusses memory considerations using the following examples:

1. *This is the malt that the rat that the cat that the dog worried killed ate.*
2. *This is the malt that was eaten by the rat that was killed by the cat that was worried by the dog.*

Whereas the first sentence is hard to understand, the second is a lot easier even if formed by the same grammar rule applied iteratively. Different hypotheses have been provided to explain this phenomenon, such as the large number of incomplete and nested syntactic relationships that must be stored in the brain. Levy shows that this phenomenon is more complex than it seems, and complexity also depends on the kind of structure used, on the arguments of the sentence, and on the distance between elements in memory.

Levy (2013) discusses expectation in terms of a traditional task in cognitive science that consists of completing the end of unfinished sentences and measuring the extent to which continuing the sentence can be predicted by humans. Let us take the two following sentences:

3. *The boat passed easily under the —*
4. *Rita slowly walked down the shaky —*

The first sentence provides a strongly predictive context (leading the speaker to generally propose “bridge” to complete the sentence) whereas the second sentence is more open, as demonstrated by the greater variety of possible answers and a longer time, on average, needed to complete the sentence.

Recently, several researchers have proposed models combining memory and expectation to measure comprehension difficulty (Boston et al. 2011). One of the main ideas is to measure complexity through large-scale probabilistic context-free grammar that can capture both dimensions of the problem (Hale 2001). These researchers have thus established a direct link between computational models and cognitive issues in language processing, providing a sound basis for empirical research. However, it should be pointed out that the results obtained are not always consistent with evidence from other kinds of studies, especially electroencephalogram or eye-tracking studies (Demberg and Keller 2008). The link between these different models deserves more attention.

1.2.2 *Language Acquisition*

One of the big puzzles in science is how children learn their native languages reliably and in a short period of time. Languages are complex systems containing large vocabularies with morphologically and derivationally inflected forms. Moreover, words from this vocabulary can be combined with a diverse inventory of syntactic constructions specific to the target language to convey some meaning in a particular context. Children have to learn to segment sounds associating forms and meanings to individual lexical items, develop a processing system to generate and comprehend sentences, and acquire pragmatic and social skills to use language in an acceptable manner in different contexts (Hyams 1987). Yet children are typically exposed to sentences that are “*propositionally simple, limited in vocabulary, slowly and carefully enunciated, repetitive, deictic, and usually referring to the here and now*” (Wanner and Gleitman 1982). So, based on this data, how can children arrive at a mature state that is so sophisticated? What are the precise mechanisms involved in acquiring a language? Are they specific to language or are they general-purpose learning mechanisms? How much do learners know about languages before exposure to a specific language? How much exposure to language is needed for successful learning? There are many questions related to language acquisition, and computational modeling has been used to address some of them (Fazly et al. 2010; Wintner 2010; Alishahi 2011; Steedman 2012; Kwiatkowski et al. 2012). Computational models usually include five components whose degree of complexity varies according to the particular focus of the research (Bertolo 2001):

- (1) The first component is a definition of what is being learned, in this case a language, and any specific subtasks, such as word segmentation

(Brent 1999; Lignos 2011; Elsner et al. 2013), morphology (Rumelhart and McClelland 1986; Legate and Yang 2007), or syntax (Berwick 1985; Briscoe 1997; Villavicencio 2002; Yang 2004; Villavicencio 2011; Kwiatkowski et al. 2012; Yang 2013).

- (2) The second component defines the available hypotheses that the learning model can formulate, that is, the hypothesis space that needs to be considered for learning (Chomsky 1965; Berwick and Niyogi 1996; Yang 2004). The trajectory of the learner in this space is driven by the input data toward the target language.
- (3) The fourth component defines the learning environment the model is exposed to. This may include the order and the frequency with which the data occur in the environment, along with any (correct or incorrect) clues about whether the data belong to the target language (Berwick and Niyogi 1996; Legate and Yang 2005; Pearl 2005).
- (4) The fourth component is a definition of the updating procedure for the learner's hypotheses along with any restrictions involved. This procedure determines how conservative the learner is in changing the current hypothesis (Niyogi and Berwick 1996; Briscoe 2000; Berwick and Niyogi 1996).
- (5) The last is a definition of success in the task. The model needs to be evaluated according to a definition of successful learning that indicates when the target language has been successfully acquired. Success criteria include those defined by such learning paradigms as identification in the limit (Gold 1967), probably approximately correct learning (Valiant 2013), and the minimum description length principle (Rissanen 1989).

One of the challenges with research in this area is that in general we have only very limited and often indirect access to the neural regions involved in language production and understanding, especially during the language acquisition period, and this is usually restricted to the output product. Corpora containing naturalistic language acquisition data from transcripts of child-directed and child-produced speech have been used as the basis for research in the area. They include data from longitudinal studies, following the same child for several years and allowing the investigation of different developmental stages. There are also latitudinal studies that include various children of particular age groups, and these may help avoid any individual bias from personal language traits. Initiatives such as CHILDES (MacWhinney 2000) have provided repositories for language acquisition data for more than twenty-five languages, some with additional information from part-of-speech taggers and parsers (Sagae et al. 2004; Villavicencio et al. 2012), others providing audio and video recordings with the transcripts.³

³ <http://childe.psych.cmu.edu>

The availability of language acquisition data brings an enormous potential for in vitro testing of different theories of acquisition via simulations in computational models (Villavicencio 2002; Perfors et al. 2010; Kwiatkowski et al. 2012; Steedman 2012). For instance, there has been much interest in hierarchical Bayesian models and their application to child language acquisition (Perfors et al. 2010; Hsu and Chater 2010; Parisien et al. 2008; Parisien and Stevenson 2010). Much of their appeal comes from being able to handle noise and ambiguity in the input data, while also accounting for known pre-existing language biases via prior probabilities. They have been applied to the acquisition of word segmentation (Pearl et al. 2010; Phillips and Pearl 2014), verb alternations (Perfors et al. 2010; Parisien and Stevenson 2010; Villavicencio et al. 2013), argument structure (Alishahi and Stevenson 2008), and multiword expressions (Nematzadeh et al. 2013), among other tasks.

1.2.3 Clinical Conditions

Because many clinical conditions have an impact on language abilities, computational methods have also been used as tools to investigate possible language changes associated with these pathologies. For example, Alzheimer's disease, which affects millions of people around the world, has from its early stages a noticeable impact on lexical search and retrieval processes. The use of computational methods that allow the creation of synthetic simulations compatible with this condition may contribute to an early diagnosis by helping to distinguish changes that are triggered by the disease from those that arise as a natural consequence of aging.

One promising line of investigation uses concepts from graph theory (Watts and Strogatz 1998) to model the lexicon as a complex network in which words or concepts correspond to nodes and are connected to one another by specific relations, such as proximity in a sentence or synonymy (Steyvers and Tenenbaum 2005; De Deyne and Storms 2008). Measures like the clustering coefficient of the network, the number of connected components, and the average length of the shortest path between pairs of nodes have been used to determine characteristics of networks in healthy and clinical cases (Cabana et al. 2011; Bertola et al. 2014), in studies related to semantic storage and the mechanisms that operate on it (De Deyne and Storms 2008; Mota et al. 2012), and in investigations that use simulations of changes that lead from healthy to clinical networks (Borge-Holthoefer et al. 2011). For example, starting from networks of semantic priming in healthy subjects Borge-Holthoefer and coworkers simulated the evolution to a clinical condition through changes in network connectivity that led to a progressive degradation of the network structure that has qualitative agreement with real observations of clinical patients with Alzheimer's disease.

Resources such as the MRC Psycholinguistic Dataset (Coltheart 1981), WordNet (Fellbaum 1998), or the University of South Florida Free Association Norms (Nelson et al. 2004) provide additional information for analyzing language data. They include such characteristics as the familiarity, concreteness, and age of acquisition of the words as well as the semantic similarity or association strength among them. However, since these resources have limited coverage and are not available for all languages, some alternatives are to also use data-driven methods to automatically extract relevant information from corpora, and to employ crowdsourcing for additional judgments (McDonald and Brew 2004; Padó and Lapata 2007; Hill et al. 2015; Köper and Schulte im Walde 2016). For instance, distributional semantic models (Lin 1998; Le and Mikolov 2014) capture semantic relatedness among words, and they have been found to successfully explain human performance in semantic priming tasks. Moreover, the more recent models based on neural networks (Mikolov et al. 2013) provided a better fit to the behavioral data (Mandera et al. *in press*).

Extensive, accurate, and fast detection of patients in early stages of pathological conditions has enormous potential for maximizing the effectiveness of treatments while minimizing their costs, and computational methods can contribute to this goal.

1.2.4 *The Origins of Language and Language Evolution*

Investigations on the origins of language have been the focus of much interest for centuries. They have examined questions that range from determining the biological mechanisms involved with language production and understanding, to how languages evolved into their contemporary variants (Bickerton 2016; Hauser et al. 2014; Berwick and Chomsky 2015). These problems have been considered some of the hardest in science (Christiansen and Kirby 2003), given the uniqueness and complexity of the human language and the lack of empirical evidence.⁴ According to Hauser et al. (2014), the way forward for empirical work involves combined efforts from:

- comparative animal behavior, looking at natural communication and artificial languages;
- paleontology and archaeology, examining structural characteristics of skulls and bones that can be linked to brain functions;
- molecular biology, mapping genes to complex behavior; and
- mathematical modeling of computations and representations.

⁴ Discussions on the topic were even banned for a while by the Linguistic Society of Paris in the 19th century (Johansson 2005).

In particular, the latter allow the definition of complex simulations involving populations of linguistic agents that interact with one another to try to approximate possible scenarios for the emergence of language. Language is seen as a complex adaptive system that may be affected by variables like the learning algorithms adopted and the communicative efficiency of competing alternatives, in addition to factors like language contact between different populations and population size. These need to be realistic models of how phonological, syntactic, and semantic representations arose and were selected for in populations, with the possibility of testing their assumptions with regards to plausibility (Hauser et al. 2014).

1.3 Content and Structure of the Book

The chapters in this collection present different aspects of research on computational models of language and cognition. They display a cross-section of recent research in the area, covering the spectrum from theoretical considerations and formalizations to more applied models and the construction of applications and resources.

The chapters in the first part of the book describe works that analyze psycholinguistic data using neural and cognitive language processing. Recordings of brain activity data are one of the most direct reflections of the states and processes involved in language processing, and when analyzed in the light of cognitive and linguistic theories they can provide insights about functions and architecture of the language faculty. Chapter 3, “Decoding Language from the Brain,” by Brian Murphy, Leila Wehbe, and Alona Fyshe, provides an overview of recent works that use computational modeling of language and of brain activity. They start with a discussion of how patterns of electrophysiological activity have been associated with sentence-processing difficulties and language complexity. For words, they describe how computational models can distinguish relevant aspects of brain activity for word meaning from noise using distributional semantic theory. Looking at syntax, they examine how lexical units are combined to form short phrases and how existing theories of language characterize the representations produced by compositional processes. They finish with a discussion of experiments that use more natural language understanding tasks for holistic and realistic language processing.

Chapter 2, entitled “Light-and-Deep Parsing: Cognitive Model of Sentence Processing” by Philippe Blache, provides an overview of language processing from various perspectives, including neurolinguistics, with findings, from electrophysiological studies. On this basis, the author argues for an alternative to the classical architectures involving modular and serial processing that takes into account language as a whole. He proposes a new representation for

linguistic information, based on properties. For basic properties that are assessable in a simple and direct manner, the default processing mechanism based on light parsing is applied. This mechanism stores words in working memory, assembles them into chunks, infers properties, and activates constructions, resulting in fast, if shallow, processing and direct access to interpretation. For more complex cases, a deeper processing needs to be adopted with classical strictly incremental and serial interpretation that is compositionally constructed from a syntactic structure.

The computational modeling of clinical groups in psycholinguistic tasks can also provide insights about language faculty by characterizing how particular conditions affect language use. In the final chapter in Part II, “Graph Theory Applied to Speech: Insights on Cognitive Deficit Diagnosis and Dream Research,” by Natalia Bezerra Mota, Mauro Copelli, and Sidarta Ribeiro, graph theory is used for structural analysis of the language used by clinical groups, represented as networks, beyond what a lexical analysis would reveal, to help in the psychiatric diagnosis of psychoses and dementias. The first study examines how networks representing the flow of thoughts of bipolar and schizophrenic patients are able to distinguish clinical from control groups based on their verbal reports of dreams or waking events. A second study looks at the use of networks representing a verbal fluency task. They model a group of clinical participants diagnosed with Alzheimer’s dementia, another with moderate cognitive impairment, and a third group of healthy elderly participants. Based on topological analysis of the networks it was possible to distinguish among the patients or subjects in these three groups.

The chapters in Part II explore the use of data-driven methods for acquiring information from large amounts of language, in tasks ranging from translation, inference about semantic roles, native language identification, and speech segmentation. Chapter 5, “Putting Linguistics Back into Computational Linguistics,” by Martin Kay, discusses the place of knowledge about languages and speakers in computational linguistics and natural language processing. For Kay communication is a collaborative task that involves the hearer guessing the speaker’s intentions. He argues that it is not enough to examine large quantities of texts to discover all we need to know about languages. The referential function of language should also be taken into account, both for a better understanding of the human language ability and for language technology. Looking at the case of translation, he analyzes the advantages of doing that comparing the syntactic and pragmatic traditions of translation. The former uses information about lexical correspondences in source and target language and possibly the reordering of words, which can be learned from huge quantities of data using statistical approaches. The latter starts from what the original author wants to communicate and finds a way of expressing it in the target language sometimes

independently of the words and phrases in the source text, and possibly making implicit information in the source explicit in the target, if important to convey the message.

In Chapter 6, “A Distributional Model of Verb-Specific Semantic Roles Inferences,” Gianluca Lebani and Alessandro Lenci start with an overview of research on acquisition and representation of thematic roles, where roles describe the relation of each of the arguments of a verb in the event or situation it expresses. Adopting the view of thematic roles as clusters of properties entailed by verb arguments, they use evidence from behavioral data to define a more fine-grained characterization of the properties activated by a verb, focusing on a subset of English verbs, and examine to what extent these properties can be acquired from corpus-based distributional data.

The next chapter in Part III is “Native Language Identification on EFCAM-DAT,” by Xiao Jiang, Yan Huang, Yufan Guo, Jeroen Geertzen, Theodora Alexopoulou, Lin Sun, and Anna Korhonen. As mentioned in its title, this chapter deals with the automatic identification of the native language of second-language learners. This has theoretical consequences, especially to determine to what extent L1 backgrounds influence L2 learning and “whether there is a significant difference between the writings of L2 learners across different L1 backgrounds.” This research domain has also immediate and practical applications, for example, in language tutoring systems and authorship profiling. The chapter offers new insights based on a new corpus, called the EF-Cambridge Open English Learner Database (EFCAMDAT), which is multiple times larger than previous L2 corpora and provides longitudinal data across sixteen proficiency levels. The system for native language identification presented in the paper employs accurate machine learning with a wide range of linguistic features. The authors report high overall accuracy of approximately 80% at low and medium proficiency levels and 70% at advanced levels, and detailed analysis shows that the top performing features differ from one proficiency level to another, which indicates that a fine-grained analysis is necessary to take into account the difference of the various learner proficiency.

Part III, “Evaluating Language Acquisition Models: A Utility-Based Look at Bayesian Segmentation,” by Lisa Pearl and Lawrence Phillips. The authors address the problem of evaluation in an unsupervised domain, especially when we have an imperfect knowledge of this domain. The problem is even more difficult when it comes to child language acquisition due to “uncertainty about the exact nature of the target linguistic knowledge and a lack of empirical evidence about children’s knowledge at specific stages in development.” The idea of a gold standard for language acquisition is thus not realistic. The rest of the paper investigates this issue through the study of initial stages of speech segmentation, in which a fluent stream of speech is divided by the learner into useful units, such as words. The authors show that segmentation based

on Bayesian models, which has proven successful for English, also obtains good results for a variety of other languages. This is particularly true if a relevant segmentation (“useful and valid nonword units”) is taken into account, which can be quite different than the traditional gold standard based on written-word segmentation. The authors conclude by showing that “this serves as a general methodological contribution about the definition of segmentation success, especially when we consider that useful units may vary across the world’s languages.”

The third and final part of the book deals with social issues in language evolution. Most people admit that the primary goal of languages is to make it possible for humans to communicate and easily exchange even complex ideas. What is not so clear is why there are so many languages around the world, and how and why these languages constantly evolve, change, and even disappear. This section provides theoretical as well as practical accounts and also considers how computational models can shed new light on this complex issue.

Chapter 9, by Anne Reboul, is “Social Evolution of Public Languages: Between Rousseau’s Eden and Hobbes’ Leviathan.” Reboul observes that nearly all models of language evolution rely on social scenarios, where language is the main tool for specific purposes like hunting, sexual selection, or tool making. Moreover, apart from when language is seen as primarily a mental tool, all hypotheses involve some social dimension. The question addressed in this chapter is whether “the social pressure leading to the emergence of language is due to prosocial attitudes” (the cooperative/altruistic hypothesis) “or to an arms race motivated by inside group competition and conflict.” The scenarios range between altruistic scenarios (Rousseauist scenario) or more conflictual ones, where competition comes before cooperation (the Hobbesian scenario). In the rest of her chapter, Reboul criticizes the idea that language is a communication system in the strong sense: language did not emerge primarily for communication. Instead, the author shows convincingly that negotiation and persuasion are more important. In this context, language is not only a tool for communication but also a perfect tool for implicit communication and argumentation. The chapter is based on recent theories of communication and argumentation and sheds new light on a hot topic in the domain.

Chapter 10, Genetic Biases in Language: Computer Models and Experimental Approaches, is by Rick Janssen and Dan Dediu. The authors observe that language evolution as an area has been highly inspired by biological models of evolution. Moreover, computational models have shown in practice how specific features can be amplified from generation to generation, leading to preferential selection of characteristic language features such as recursion, compositionality, and other universal features. In these models the evolution of languages is based on specific biological characteristics of the human species, encoded in the human genome, but agents might evolve to a state of predisposed

adaptability, whereas “culturally stable language features might get assimilated into the genome via Baldwinian niche construction.” Although this issue is largely controversial, it may be considered a valid alternative to the adaptation of language-specific features, “for example explaining speech perception as a possible co-option of more general learning and pattern recognition mechanisms.” The authors claim that the evolution of language cannot be explained solely from a biological perspective and that social interaction must also be taken into account. Computational and agent-based models give a sound basis for this thesis that deserves to be exposed and discussed among researchers.

The last chapter in this section, by Remi Van Trijp is “Transparency versus Processing Efficiency: A Case Study on German Declension.” The author addresses ambiguity in natural languages and argues that it may lead to greater efficiency in language processing. The claim is supported by a case study on the German declension system. Van Trijp proposes a formalization that shows case syncretism is “efficiently processed as long as the case forms are still in functional opposition of each other.” Syncretism and ambiguity should thus be studied within the whole linguistic system or “linguistic ecosystem” according to the author.

Acknowledgments

Thierry Poibeau was partly sponsored by TransferS (laboratoire d'excellence, program Investissements d'avenir ANR-10-IDEX-0001-02 PSL* and ANR-10-LABX-0099).

Aline Villavicencio was partly sponsored by projects AIM-WEST (FAPERGS-INRIA 1706- 2551/13-7), CNPq 423843/2016-8, 312114/2015-0, “Simplificação Textual de Expressões Complexas,” sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

We the editors thank all the contributors for working with us to produce this book, and all the reviewers for their insightful comments.

References

- Abney, Steven. 1991. Parsing by Chunks. In: Berwick, Robert, Abney, Steven, and Tenny, Carol (eds), *Principle-Based Parsing*. Dordrecht: Kluwer Academic Press.
- Alishahi, Afra. 2010. Computational modeling of human language acquisition. *Synthesis Lectures on Human Language Technologies*, 3(1), 1–107.
- Alishahi, Afra, and Stevenson, Suzanne. 2008. A computational model of early argument structure acquisition. *Cognitive Science*, 32(5), 789–834.
- Allen, James. 1994. *Natural Language Understanding*. New York: Pearson.

- Anderson, John R. 1976. *Language, Memory, and Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Anderson, John R., Bothell, Dan, Byrne, Michael D., Douglass, Scott, Lebiere, Christian, and Qin, Yulin. 2004. An integrated theory of the mind. *Psychological Review*, **111**(4), 1036–1060.
- Bergen, Benjamin, and Chang, Nancy. 2003. Embodied Construction Grammar in Simulation-Based Language Understanding. In: Ostman, J.-O., and Fried, M. (eds), *Construction Grammar(s): Cognitive and Cross-Language Dimensions*. Amsterdam: Johns Benjamins.
- Bertolo, Laiss, Mota, Natalia Bezerra, Copelli, Mauro, Rivero, Thiago, Satler, Breno, Diniz, De Oliveira, Romano, Marco Aurelio, Ribeiro, Sidarta, and Malloy-diniz, Leandro Fernandes. 2014. Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in Aging Neuroscience*, **6**(185).
- Bertolo, Stefano. 2001. A Brief Overview of Learnability. Pages 1–14 of: Bertolo, Stefano (ed), *Language Acquisition and Learnability*. Cambridge University Press.
- Berwick, Robert C. 1985. *The Acquisition of Syntactic Knowledge*. MIT Press.
- Berwick, Robert C., and Chomsky, Noam. 2015. *Why Only Us: Language and Evolution*. Cambridge, MA: MIT Press.
- Berwick, Robert C., and Niyogi, Partha. 1996. Learning from Triggers. *Linguistic Inquiry*, **27**(4), 605–622.
- Bickerton, Derek. 2016. *Roots of Language*. Berlin: Language Science Press.
- Blache, Philippe. 2013. Chunks et activation: Un modèle de facilitation du traitement linguistique. In: *Proceedings of the Conference Traitement Automatique du Langage Naturel*. Les Sables d'Olonne: ATALA.
- Borge-Holthoefer, Javier, Moreno, Yamir, and Arenas, Alex. 2011. Modeling abnormal priming in Alzheimer's patients with a free association network. *PLoS ONE*, **6**(8).
- Boston, Marisa F., Hale, John T., Vasishth, Shravan, and Kliegl, Reinhold. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, **26**(3), 301–349.
- Brachman, Ronald, and Levesque, Hector. 2004. *Knowledge Representation and Reasoning*. San Francisco: Morgan Kaufmann Publishers Inc.
- Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, **34**(1), 71–105.
- Briscoe, Ted. 1997. Co-evolution of language and of the language acquisition device. Pages 418–427 of: *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Briscoe, Ted. 2000. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 245–296.
- Cabana, Álvaro, Valle-lisboa, Juan C., Elvevåg, Brita, and Mizraji, Eduardo. 2011. Detecting order-disorder transitions in discourse: Implications for schizophrenia. *Schizophrenia Research*, **131**(1–3), 157–164.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Christiansen, Morten H., and Kirby, Simon. 2003. Language Evolution: The Hardest Problem in Science? In: Christiansen, M.H., and Kirby, S. (eds), *Language Evolution: The States of the Art*. New York: Oxford University Press.

- Chrupala, Grzegorz, Ákos Kádár, and Alishahi, Afra. 2015. Learning language through pictures. Pages 112–118 of: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*. Beijing: ACL.
- Coltheart, Max. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33A**, 497–505.
- De Deyne, Simon, and Storms, Gert. 2008. Word associations: Network and semantic properties. *Behavior research methods*, **40**(1), 213–231.
- Demberg, Vera, and Keller, Frank. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, **109**(2), 193–210.
- Elsner, Micha, Goldwater, Sharon, Feldman, Naomi, and Wood, Frank. 2013. A Joint Learning Model of Word Segmentation, Lexical Acquisition, and Phonetic Variability. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Erk, Katrin, and McCarthy, Diana. 2009. Graded word sense assignment. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference*. ACL.
- Fazly, Afsaneh, Alishahi, Afra, and Stevenson, Suzanne. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, **34**(6), 1017–1063.
- Fellbaum, Christiane (ed). 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication Series. Cambridge, MA: MIT Press.
- Ferrucci, David A. 2012. Introduction to “This is Watson.” *IBM J. Res. Dev.*, **56**(3), 235–249.
- Francez, Nissim, and Wintner, Shuly. 2012. *Unification Grammars*. Cambridge, UK: Cambridge University Press.
- Fu, Ruiji, Guo, Jiang, Qin, Bing, Che, Wanxiang, Wang, Haifeng, and Liu, Ting. 2014. Learning Semantic Hierarchies via Word Embeddings. Pages 1199–1209 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL 2014.
- Gee, James P., and Grosjean, François. 2004. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, **15**(4), 411–458.
- Gold, E. Mark. 1967. Language identification in the limit. *Information and Control*, **10**(5), 447–474.
- Hale, John T. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In: *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Hauser, Marc D, Yang, Charles, Berwick, Robert C, Tattersall, Ian, Ryan, Michael J, Watumull, Jeffrey, Chomsky, Noam, and Lewontin, Richard C. 2014. The mystery of language evolution. *Frontiers in Psychology*, **5**(May), 1–12.
- Hill, Felix, Reichart, Roi, and Korhonen, Anna. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, **41**(4), 665–695.
- Holland, John H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. Cambridge, MA: MIT Press.
- Hsu, Anne S., and Chater, Nick. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science*, **34**(6), 972–1016.

- Hyams, Nina. 1987. The theory of parameters and syntactic development. Pages 1–22 of: Roeper, Thomas, and Williams, Edwin (eds), *Parameter Setting*. Dordrecht: Springer Netherlands.
- Johansson, Sverker. 2005. *Origins of Language: Constraints on Hypotheses*. Converging Evidence in Language and Communication Research. Amsterdam: John Benjamins Publishing Company.
- Joshi, Aravind K. 1990. Processing crossed and nested dependencies: An automaton perspective on the psycholinguistic results. *Language and Cognitive Processes*, **5**(1), 1–27.
- Jurafsky, Daniel, and Martin, James H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River: Pearson Prentice Hall.
- Köper, Maximilian, and Schulte im Walde, Sabine. 2016. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. Pages 2595–2598 of: *Proceedings of the 10th International Conference on Language Resources and Evaluation*.
- Kwiatkowski, Tom, Goldwater, Sharon, Zettlemoyer, Luke, and Steedman, Mark. 2012. A Probabilistic Model of Syntactic and Semantic Acquisition from Child-directed Utterances and Their Meanings. Pages 234–244 of: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12.
- Le, Quoc, and Mikolov, Tomas. 2014. Distributed Representations of Sentences and Documents. In: *31st International Conference on Machine Learning*.
- Legate, Julie Anne, and Yang, Charles. 2005. The richness of the poverty of the stimulus. *On the Occasion of Happy Golden Anniversary, Generative Syntax: 50 Years Since Logical Structure of Linguistic Theory*.
- Legate, Julie Anne, and Yang, Charles. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition*, **14**(3), 315–344.
- Lenat, Douglas B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, **38**(11), 33–38.
- Levy, Roger. 2013. Memory and surprisal in human sentence comprehension. Page 78114 of: van Gompel, Roger P. G. (ed), *Sentence Processing*. Hove, UK: Psychology Press.
- Lignos, Constantine. 2011. Modeling Infant Word Segmentation. Pages 29–38 of: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*.
- Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. Pages 768–774 of: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.
- Liu, Hugo, and Singh, Push. 2004. ConceptNet – A practical commonsense reasoning tool-kit. *BT Technology Journal*, **22**(4), 211–226.
- MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mandera, Paweł, Keuleers, Emmanuel, and Brysbaert, Marc. in press. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

- Manning, Christopher D., and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- McDonald, Scott, and Brew, Chris. 2004. A Distributional Model of Semantic Context Effects in Lexical Processing. Pages 17–24 of: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.
- Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mitchell, Thomas M. 1997. *Machine Learning*. New York: McGraw-Hill.
- Moor, James H. 2003. Turing test. Pages 1801–1802 of: *Encyclopedia of Computer Science*. Chichester, UK: John Wiley and Sons.
- Mota, Natalia B, Vasconcelos, Nivaldo A P, Lemos, Nathalia, Pieretti, Ana C, Kinouchi, Osame, Cecchi, Guillermo A, Copelli, Mauro, and Ribeiro, Sidarta. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *7*(4), 1–9.
- Nelson, Douglas L., McEvoy, Cathy L., and Schreiber, Thomas A. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, **36**(3), 402–407.
- Nematzadeh, Aida, Fazly, Afsaneh, and Stevenson, Suzanne. 2013. Child acquisition of multiword verbs: A computational investigation. Pages 235–256 of: Villavicencio, A., Poibeau, T., Korhonen, A., and Alishahi, A. (eds), *Cognitive Aspects of Computational Language Acquisition*. Berlin: Springer.
- Nirenburg, Sergei, and Raskin, Victor. 2004. *Ontological Semantics*. Cambridge, MA: MIT Press.
- Niyogi, Partha, and Berwick, Robert C. 1996. A language learning model for finite parameter spaces. *Cognition*, **61**(1-2), 161–193.
- Padó, Sebastian, and Lapata, Mirella. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.
- Pantel, Patrick, and Lin, Dekang. 2002. Discovering Word Senses from Text. Pages 613–619 of: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- Parisien, Christopher, and Stevenson, Suzanne. 2010. Learning verb alternations in a usage-based Bayesian model. In: *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Parisien, Christopher, Fazly, Afsaneh, and Stevenson, Suzanne. 2008. An incremental Bayesian model for learning syntactic categories. Pages 89–96 of: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*.
- Pearl, Lisa. 2005. The Input for Syntactic Acquisition: Solutions from Language Change Modeling. Pages 1–9 of: *Proceedings of the Workshop on Psychocomputational Models of Human Language Acquisition*.
- Pearl, Lisa, Goldwater, Sharon, and Steyvers, Mark. 2010. How ideal are we? Incorporating human limitations into Bayesian models of word segmentation. Page 315–326 of: *Proceedings of the 34th Annual Boston University Conference on Child Language Development*. Somerville, MA: Cascadilla Press.
- Perfors, Amy, Tenenbaum, Joshua B, and Wonnacott, Elizabeth. 2010. Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language*, **37**(03), 607–642.

- Phillips, Lawrence, and Pearl, Lisa. 2014. Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In: *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.
- Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry*. Series in Computer Science, vol. 15. World Scientific.
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, *Psychological and Biological Models*. Cambridge, MA: MIT Press.
- Sagae, Kenji, MacWhinney, Brian, and Lavie, Alon. 2004. Adding Syntactic Annotations to Transcripts of Parent-Child Dialogs. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*.
- Sowa, John (ed). 1991. *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann Publishers.
- Steedman, Mark. 2012. Probabilistic Models of Grammar Acquisition. Pages 19–29 of: *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*.
- Steyvers, Mark, and Tenenbaum, Joshua B. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, **29**, 41–78.
- Tiedemann, Jörg. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Turing, Alan. 1950. Computing machinery and intelligence. *Mind*, **236**, 433–460.
- Turney, Peter D., and Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Int. Res.*, **37**(1), 141–188.
- Valiant, Leslie. 2013. *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. New York: Basic Books, Inc.
- Villavicencio, Aline. 2002. The acquisition of a unification-based generalised categorial grammar. PhD thesis, University of Cambridge.
- Villavicencio, Aline. 2011. Language acquisition with a unification-based grammar. In: K. Borjars, R. Borsley (ed), *Non-transformational Syntax: Formal and Explicit Models of Grammar*. Chichester, West Sussex, UK; Malden, MA: Wiley-Blackwell.
- Villavicencio, Aline, Yankama, Beracah, Wilkens, Rodrigo, Idiart, Marco, and Berwick, Robert. 2012. An annotated English child language database. Pages 23–25 of: *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*.
- Villavicencio, Aline, Idiart, Marco, Berwick, Robert, and Malioutov, Igor. 2013. Language Acquisition and Probabilistic Models: Keeping It Simple. Pages 1321–1330 of: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Wanner, Eric, and Gleitman, Lila R. (eds). 1982. *Language Acquisition: The State of the Art*. Cambridge, UK: Cambridge University Press.
- Watts, Duncan J., and Strogatz, Steven H. 1998. Collective dynamics of ‘small-world’ networks. *Nature*, **393**(June), 440–442.
- Weizenbaum, Joseph. 1966. ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM*, **9**(1), 36–45.

- Wintner, Shuly. 2010. Computational Models of Language Acquisition. Pages 86–99 of: Gelbukh, Alexander (ed), *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010*.
- Yang, Charles. 2004. Universal grammar, statistics or both? *Trends in Cognitive Sciences*, **8**(10), 451–456.
- Yang, Charles. 2013. Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*, **110**(16), 6324–6327.

Part II

Models of Neural and Cognitive Processing

2 Light and Deep Parsing: A Cognitive Model of Sentence Processing

Philippe Blache

Abstract

Humans process language quickly and efficiently, despite the complexity of the task. However, classical language-processing models do not account well for this feature. In particular, most of them are based on an incremental organization, in which the process is homogeneous and consists in building step-by-step a precise syntactic structure, from which an interpretation is calculated. In this chapter, we present evidence that contradicts this view, and show that language processing can be achieved at varying levels of precision. Often, processing remains shallow, leaving interpretation greatly underspecified.

We propose a new language-processing architecture, involving two types of mechanisms. We show that, in most cases, shallow processing is sufficient and deep parsing is required only when faced with difficulty. The architecture we propose is based on an interdisciplinary perspective in which elements from linguistics, natural language processing, and psycholinguistics come into play.

2.1 Introduction

How humans process language quickly and efficiently remains largely unexplained. The main difficulty is that, although many disciplines (linguistics, psychology, computer science, and neuroscience) have addressed this question, it is difficult to describe language as a global system. Typically, no linguistic theory entirely explains how the different sources of linguistic information interact. Most theories, and then most descriptions, only capture partial phenomena, without providing a general framework bringing together prosody, pragmatics, syntax, semantics, etc. For this reason, many linguistic theories still consider language organization as modular: linguistic domains are studied and processed separately, their interaction is implemented at a later stage. As a consequence, the lack of a general theory of language, accounting for its different aspects, renders difficult the elaboration of a global processing architecture. This problem

has direct consequences for natural language processing: the classical architecture relies on different subtasks: segmenting, labeling, identifying the structures, interpreting, etc. This organization more or less strictly imposes a sequential view of language processing, considering in particular words as being the core of the system. Such a view does not account for the fact that language is based on complex objects, made of different and heterogeneous sources of information, interconnected at different levels, and which interpretation cannot always be done compositionally (each information domain transferring a subset of information to another).

Cognitive approaches to language processing (LP) face the same difficulties. Even more crucially than for linguistics, psycholinguistics models mainly rely on a sequential and modular organization. Language is usually considered to be strictly incremental, relying on a word-by-word processing, each word being integrated into a partial syntactic structure, starting from which an interpretation can be calculated. In this organization, the different steps used in classical natural language processing (NLP) architectures are implemented: segmentation, lexical access, categorization, parsing, interpretation, etc.

This perspective is also adopted in neurolinguistics, trying to identify in a spatial or in a temporal dimension the brain basis of LP. The question there consists in studying what parts of the brain are involved in LP and in what manner. What is interesting is that even though the different works focus on only one linguistic dimension (e.g., lexicon, lexical semantics, prosody, morphosyntax), they also show that they are strongly dependent on each other.

We propose in this paper an approach bringing closer the different types of knowledge about LP coming from these disciplines that makes it possible to draw a broader and more integrated architecture.

2.2 An Interdisciplinary View of Language Processing

This section gives an overview of language processing through different disciplines (linguistics, psycholinguistics, computational linguistics, and neurolinguistics). We show in particular that classical architectures (modular and serial) are now challenged, in particular when taking into account language as a whole, in its natural environment, opening the door to a more flexible approach.

2.2.1 A Classical View: Modular and Incremental LP

This section presents the main features of the classical modular architecture of LP, from the generative framework, that has influenced other disciplines.

Modularity: A view from linguistics: The classical generative architecture relies on a succession of different modules, each one specialized for a

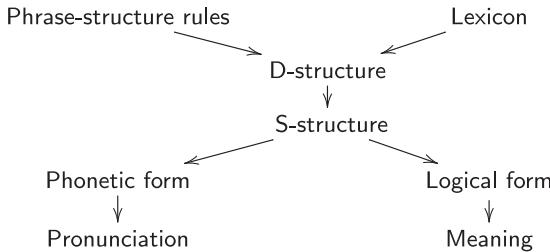


Figure 2.1 A classical generative architecture of language processing.

specific linguistic dimension. The Figure 2.1 illustrates such organization for the “Government and Binding” approach [Chomsky, 1981]. Starting from the lexicon and the rules (the set of local trees), the process consists of generating an underlying structure, subject to modifications and transformations that lead to another structure, closer to the surface form. From this structure, the phonological and logical forms are produced, making it possible to access the meaning:

This organization considers the different modules as not only separate but also sequential. Many linguistic theories propose a similar organization in which each domain produces a structure that is transferred to another one. One of the reasons for this is that linguistic theories are usually *syntactocentric*: all domains are considered in terms of their relation to the syntactic structure.

Module interaction: A view from computational linguistics: LP is taken from a specific perspective in NLP because of implementation constraints: LP is usually considered as a set of tasks, implementing the different modules in a serial manner. In this architecture, modules are synchronized, the input of one module being the output of the previous one. Up to now, no real answer to the question of the integration of the different sources of linguistic knowledge is given and their interaction is described in terms of specific synchronization rules [Jackendoff, 2007]. More precisely, even though many works have been done concerning the study of the interaction between the domains (e.g., prosody/syntax, syntax/semantics, etc.), solutions are proposed by giving the priority to one domain, usually syntax. For example, the compositional view of semantics [Werning et al., 2012] is implemented by the construction of a syntactic structure starting from which the interpretation can be calculated [Copestake et al., 2001]. The same kind of approach can be found for the prosody/syntax interface, in which prosodic information is integrated to the syntactic structure [Steedman, 2000]. As presented in the previous section, Modularity: A View from Linguistics, this is a syntactocentric organization, which induces an incremental and modular view of LP. As a consequence,

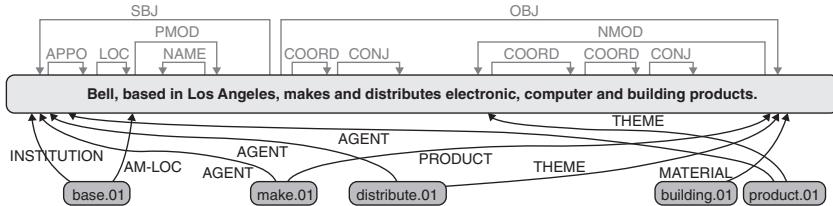


Figure 2.2 Output of the Stanford parser.

classical NLP architectures are organized around a series of subprocesses: tokenization, tagging, parsing, discourse organization, semantic interpretation. A parser builds complex structures (trees, feature-value structures, etc.) involving information at different levels, as shown in Figure 2.2, based on information produced by parsers such as the Stanford [de Marneffe and Manning, 2008], enriched with semantic information.

Other processes can be added to this general schema when studying audio (phonetics, prosody) or multimodality (gestures). Even though the question of parallelization in NLP is regularly addressed [Adriaens and Hahn, 1994; Jindal et al., 2013], the answer is usually given in terms of different parallel processes with meeting points, instead of an integrated view.

Incrementality: A view from computational psycholinguistics: In psycholinguistics, the LP classical architecture relies on the idea that processing is incremental, consisting in integrating each new word into a partial structure under construction [Fodor and Ferreira, 1998; Grodner and Gibson, 2005; Sturt and Lombardo, 2005; Keller, 2010; Altmann and Mirković, 2009; Rayner and Clifton, 2009]. In this approach, the basic units are considered to be the words: all information related to the lexical item is accessed when encountering a new word and is used to integrate the item into a partial syntactic structure (often called the *current partial phrase marker*). This operation consists in finding a site in the structure to which to attach the word. If this becomes difficult, the word is integrated where it least severely violates the grammar, following the “attach anyway” principle proposed by Fodor and Inoue [Fodor and Inoue, 1998].

This concept is also, syntactocentric, organizing information around the syntactic structure. Moreover, it is essentially sequential, in the sense that lexical information is processed before syntax, from which interpretation becomes possible. In other words, it is a *modular syntax-first* concept, supported by several classical works [Fodor, 1983; Frazier and Fodor, 1978] and still at work in many psycholinguistics models.

In terms of interpretation or meaning access, these approaches are also basically compositional, whether they are serial or parallel [Gibson, 2000]. In serial

models, the language processor initially computes only one of the possible interpretations [Fodor and Ferreira, 1998; Gorrell, 1995]. When this interpretation becomes difficult or even impossible, another interpretation is built. In parallel models, all possible interpretations are computed at once, the analysis with the greatest support being chosen over its competitors [MacDonald et al., 1994; Marslen-Wilson and Tyler, 1980; Spivey and Tanenhaus, 1998]. These two options both rely on an incremental view: interpretation is built at each new word, on the basis of a word-by-word syntactic and semantic analysis. Many issues are raised with these models. First, they both consider incrementality in a strict manner: an interpretation covering all the words at a given position is built, even if it builds an ill-formed structure [Fodor and Ferreira, 1998]. Moreover, the question of memory remains an issue in both cases: What elements are to be stored, under what form, requiring what capacity?

Brain basis of a modular architecture: A view from neurolinguistics:

The study of the physiological and brain basis of language processing also leads to different LP architecture. Among the possible investigation techniques for the exploration of LP neural correlates, electrophysiological studies are frequently used. These experiments focus on the study of event-related potentials (ERPs), which are potential changes measurable from the scalp and which reflect the activity of a specific neural process [Luck, 2005]. LP modulates a number of ERP components, located between 100 and 600 ms after the stimulus (for example, reading a word). Figure 2.3 shows the main positive (P1, P2, P3) and negative (N1, N2) deflections that can be elicited by language processing.

Many effects have been explored, related to different linguistic domains (prosody, morphology, syntax, semantics in particular) [Kutas et al., 2006; Kaan, 2007]. Even though no electric component is strictly related with one domain, we can find in the literature. Although no electrical component relates

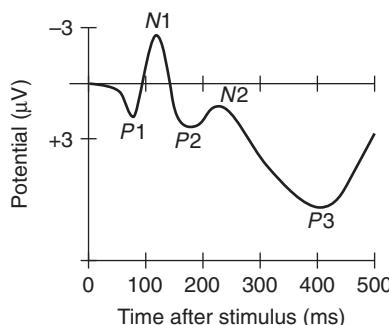


Figure 2.3 The main ERP components in language processing.

strictly to one domain, some early effects are reported for speech perception at 100 ms after the stimulus, word production at approximately 200 ms, semantics at approximately 400 ms, and syntax at approximately 600 ms. This is only a very rough picture, and all the observed effect depends on the linguistic material, in particular the amount of information coming from each domain.

Several works in neurolinguistics support a modular and serial view of language processing. Typically, the three-phases model [Friederici, 2002; Friederici, 2011] proposes an organization into three different steps, after an initial phase of acoustic-phonological analysis:

- Phase 1: Local phrase structure is built on the basis of word category information.
- Phase 2: Syntactic and semantic relations (verb/argument, thematic role assignment).
- Phase 3: Integration of the different information types and interpretation.

This organization can be completed, in an auditory comprehension model, by adding interaction of prosody at each of these stages. Different language ERP components are in relation with these phases (see Figure 2.4).

- Early left anterior negativity (ELAN, 120–200 ms): Initial syntactic structure-building processes.
- Centroparietal negativity (N400, 300–500 ms): Semantic processes.

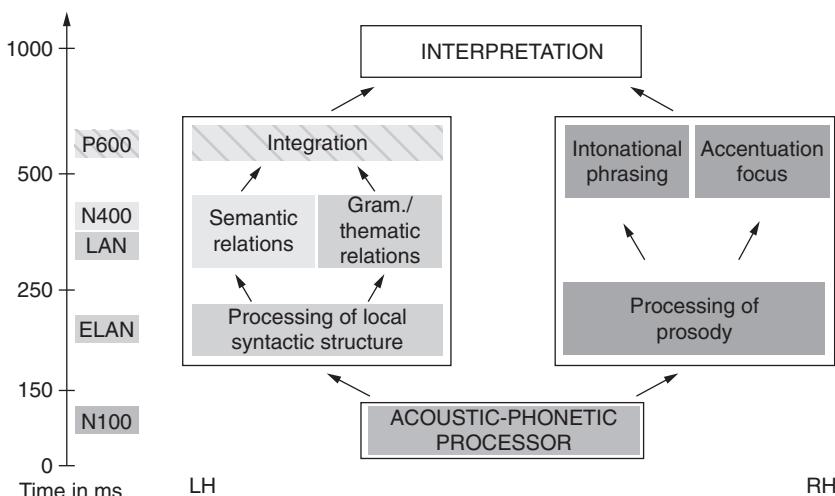


Figure 2.4 General organization of the three-phases model [Friederici, 2011].
 ELAN, early left anterior negativity; LAN, left anterior negativity; N100, _____; N400, centroparietal negativity; P600, late centroparietal positivity.

- Left anterior negativity (LAN, 300–500 ms): Grammatical relation between arguments and verb, assignment of thematic relations.
- Late centroparietal positivity (P600): Late syntactic processes.

This model is syntax-first, and consists in building a syntactic structure from which interpretation can be done.

2.2.2 An Integrated View of LP: Constructions

The classical modular view is now challenged by recent theories considering that no specific domain is at the center of the architecture, and the processing is described in terms of interaction between them. Instead of being serial, processes are considered parallel, as described in [Jackendoff, 2007].

Construction grammar [Fillmore, 1988; Goldberg, 1995] is one of those theories proposing an alternative organization. Here, no structure is predefined, and no domain need to be described before another; all linguistic phenomena are described thanks to a set of interacting properties. As presented in Goldberg, 2003, constructions are *form and meaning* pairings: a set of properties makes it possible to characterize a construction, the meaning of which is accessed directly. Constructions can be of different types, as presented in the following examples:

- Ditransitive construction: Subj V Obj1 Obj2: *She gave him a kiss*.
- Covariational conditional construction: The Xer the Yer: *The more I read the less I understand*.
- Idiomatic constructions: *Kick the bucket*; *to put all eggs in one basket*.

What is important with constructions is the fact that they are defined on the basis of different properties, possibly coming from different linguistic domains, without requiring preliminary complete analysis of each of these domains. For example, a syntactic tree is of no use in understanding an idiomatic construction. In such cases, instead of being built compositionally, the meaning of the construction is accessed directly.

This means that two types of mechanisms coexist in LP: one based on a compositional architecture, and another relying on direct access. In the first case, the architecture consists in analyzing all sources of information and their interactions. Each source or combination of sources contains a partial meaning, and their composition leads to a complete interpretation of the message. In the direct access case, the different properties, instead of bearing part of the meaning, play the role of cues in identifying the construction. The recognition of such pattern leads to a direct interpretation, without any composition. In some cases, only a few properties makes it possible to recognize and to interpret an entire construction.

2.2.3 Different Levels of Processing: LP Is Often Shallow

A flexible model of LP, the *good-enough theory* [Ferreira and Patson, 2007], has been proposed. It is based on the observation that interpretation of complex material is often shallow and incomplete. For example, Swets and colleagues [Swets et al., 2008] showed in a self-paced reading study that when participants expect superficial comprehension questions, ambiguous sentences are read faster, showing that no precise attachment resolution is done, leading to underspecified semantic representations. In this case, it is suggested that the ambiguity is not resolved, explaining the facilitation effect.

Several experiments confirm this observation that sentence comprehension can be quite shallow. For example, thematic role assignment can be subject to a simple heuristic: the first NP is the agent, the second the entity affected by the action. The use of such a heuristic has been exhibited by simple experiments showing that the interpretation of sentences contradicting this heuristic leads more often to misinterpretations than those satisfying it. These observations tend to show that, in several cases, no compositional processing is at work. Instead, as Ferreira and Patson explained, “the comprehension system tries to construct interpretations over small numbers of adjacent words whenever possible and can be lazy about computing a more global structure and meaning.” The building of a complete and precise interpretation is often delayed or even never done, replaced by the identification of “*islands*,” from which a general interpretation can be approximated.

Note that this theory contradicts several classical language-processing models. In this case, there is no systematic instantiation of thematic roles, at least in a first stage, contrary to what is required in generative theories. This is contradicts some psycholinguistic difficulty models [Gibson, 2000], which stipulate that NP without thematic roles (as well as unassigned thematic roles) impose a burden on working memory. Conversely, the good-enough theory proposes that shallow semantic processing, even involving such underspecification, can be an element of facilitation.

Several works in neurolinguistics focusing on semantic processing also suggest that some type of basic and shallow processing can be at play. In line with the good-enough theory, meaning integration can be switched off when the context renders it unnecessary. This is the case when processing idioms: semantic composition might be not fully engaged during comprehension [Rommers et al., 2013]. In particular, the activation of literal word meanings is only carried out when necessary. Analyzing the cortical responses of semantic violations (see Figure 2.5) shows no significant difference at N400 (the component related to semantic surprisal) between hard and soft violations for idiomatic context, whereas an important reduction in N400 amplitude appears for soft violation compared with hard violation for literal contexts.

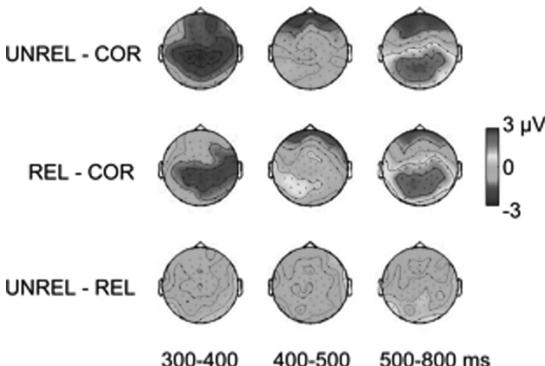


Figure 2.5 The differences between conditions in idiomatic context (COR, correct sentence; REL, soft violation; UNREL, hard violation). The schema shows the difference in the potentials when compared the correct and both types of violation, but no difference between hard and soft violations (UNREL-REL), from [Rommers et al., 2013].

2.2.4 Basic Processing Units: Words or Chunks?

One important question is that of the types of units that are used during LP. The classical option, considering LP as strictly incremental and compositional, consists in recognizing atomic elements and aggregating them progressively (phonemes, morphemes, words, phrases, etc.). This mechanism leads to an interpretation built by composition of the semantic information available at each level (in particular words and phrases). An alternative option considers that the data input stream (heard or read) is stored in the working memory on the basis of larger units, made of sets of words (also called *chunks*), grouped thanks to a shallow processing, which becomes the basis of the interpretation. Several experimental observations support this idea of a more global not strictly incremental processing.

Many NLP applications are based on such low-level information, relying on the identification of basic relationships between words (co-occurrence, order). This technique, *shallow parsing* [Uszkoreit, 2002; Balfourier et al., 2002; Baldwin et al., 2002], leads to the construction of chunks [Abney, 1991] that consist of groups of adjacent words, usually identified on the basis of their boundary markers rather than the syntactic relationships between their constituents:

[When I read] [a sentence], [I read it] [a chunk] [at a time]

Several works have shown that chunks can be considered a relevant basic unit for LP. For example, studying eye movements when reading a text shows

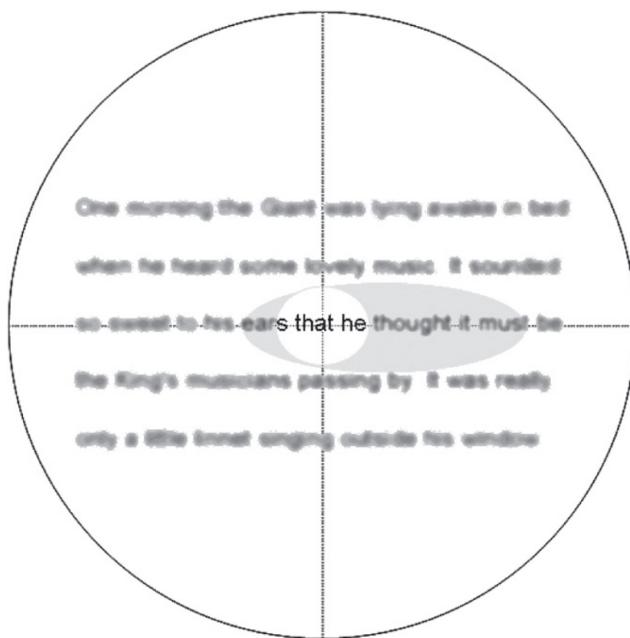


Figure 2.6 Parafoveal vision. Extracting features from the surrounding words.

that fixations are only done from time to time. This is very well known and is the result of *parafoveal vision*, which consists of a preview of adjacent words. As shown in Figure 2.6 (from Schuett et al., 2008), readers extract during a fixation visual information from the foveal visual field (central white oval) and the parafoveal visual field (grey ellipse).

This process makes it possible to extract information about upcoming words, opening the capacity to deal with entire sequences, not only separate words. Moreover, Rauzy and Blache [2012] showed that fixation can be done in chunks (defined here by a sequence of function word and content word).

This observation is an argument in favor of a more global treatment, including at the physiological level. It can be supported by other observations focusing on the neural correlates of LP: several works have specifically studied the question of syntax and, more precisely, its role in the processing of basic properties. In particular, some morphosyntactic properties can be assessed automatically, at a low level, when studying differences between chunks with or without *Det-N* or *Pro-V* agreement violations [Pulvermüller et al., 2008; Pulvermüller, 2010]. When comparing the two conditions, one observes a difference in the cortical reaction at a very early stage (around 100 ms after the stimulus, see Figure 2.7).

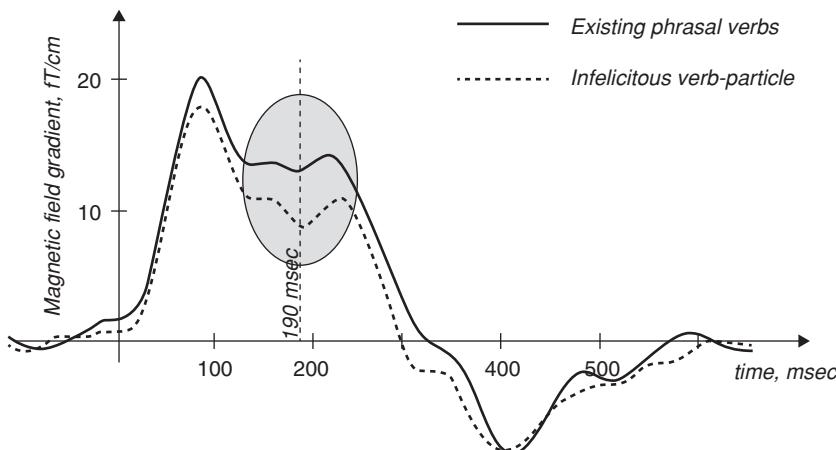


Figure 2.7 Early EEG effects of syntactic violation (mismatch negativity), from [Cappelle et al., 2010].

This effect, called *mismatch negativity*, occurs in a time range and experimental design in which there is no strictly conscious activity. This research suggests that syntax can function as a discrete combinatorial system implemented by discrete combinatorial neuronal assemblies in the brain [Pulvermüller, 2010] that can connect categories into larger constructions.

2.2.5 Intermediate Summary

Many of the observations presented to this point tend to indicate that the organization of language processing is not homogeneous. The classical, but also simplest way to explain LP consists in distinguishing different modules that are organized in a serial schema. Supporting this idea, many experiments (judgments, eye movement, brain activity) have shown the respective effects and contribution to these modules in global processing. They usually rely on a theoretical concept of language in which each module bears a subpart of the information, the global interpretation resulting from their composition.

However, this approach appears to be too simple when faced with the description of natural data. First, we know that language is intrinsically heterogeneous, made of different sources of information, different modalities, that interact at any time, producing a complex signal. In a natural environment (typically conversation), the linguistic signal is made of different sources that cannot be strictly separated and analyzed independently from the others. It is made of multiple streams that are not strictly temporally aligned.

This view of language fits better with many observations presented here so far. In particular, language processing often stays at a shallow level, leading to incomplete processing: chunks, identified in terms of basic properties instead of complete analysis, can be considered the basic processing unit, and give access to a certain level of meaning and interpretation.

These objects can be at different levels and result from the convergence of different sources of information. We distinguish between chunk and construction as follows. A *chunk* is a group of words that are gathered on the basis of low-level morphosyntactic properties. A *construction* is a chunk or a set of chunks that can be associated with a global meaning (which can be figurative).

Chunks and constructions are described as sets of interacting properties instead of structures that are built step-by-step from atomic to complex objects. These properties can be at a low level and are automatically assessed.

At the interpretation level, in line with the notion of construction, we have seen that meaning can in some cases be accessed directly instead of compositionally (e.g., idioms, multiword expressions). Moreover, interpretation is often incomplete or underspecified. In particular, it has been shown that ambiguity can be left unresolved and interpretation delayed (or even never completely built).

We propose to take into consideration these different features, gathering them into a language-processing architecture capable of accounting for different types of processing, at different levels, depending on the type of input available. The objective is to describe any type of situation, from the more controlled (e.g., laboratory speech, isolated words) to the more natural (that is, conversation). The proposal relies on the idea that, according to the context and the sources of information, LP can be either serial, modular, and compositional or, conversely, parallel, integrated, and directly interpretable. This approach induces a hybrid processing: one, at a low level, is shallow and partial and supposed to operate by default. The second, which relies on deep, modular, and compositional parsing, is activated when processing complex material (in other words, when interpretation becomes difficult). This organization comes with several assumptions:

- Instead of a word-by-word parsing, LP is based on chunks.
- Chunks are specified in terms of low-level properties, automatically assessed (i.e., without needing deep analysis).
- Semantic interpretation can be delayed.
- Chunks offer the possibility of direct access to the meaning.

In the remainder of this chapter, we will investigate these aspects by addressing specific questions:

- What is the nature of basic properties?: constraints.
- How can basic properties specify entire chunks?: constraint interaction.
- How to access directly from low-level properties to meaning: constructions.

2.3 The Theoretical Framework: Property Grammars

We present in this section the main features of *property grammar* (PG) [Blache, 2000]. PG is a linguistic theory that proposes a constraint-based processing architecture. More precisely, all linguistic information in PG is represented by means of different properties (implemented as constraints). At the difference with the classical generative paradigm, there is no specific module: all properties are mutually independent, offering the possibility to represent separately the different types of information, whatever their domain (morphology, syntax, semantics, etc.) or their level (relationships between features, categories, chunks, etc.). These properties connect the different words of a sentence when processing an input. As a consequence, instead of building a structure, the processing mechanism consists here in describing the input by identifying its different properties. Focusing on syntax and semantics, the following list summarizes the possible relationships between words:

- *Linearity*: Linear order that exists between two words.
- *Co-occurrence*: Mandatory co-occurrence between two words.
- *Exclusion*: Impossible co-occurrence between two words.
- *Uniqueness*: Impossible repetition of a same category.
- *Dependency*: Syntactic-semantic dependency between two words. Different types of dependencies are encoded: complement, subject, modification, specification, etc.

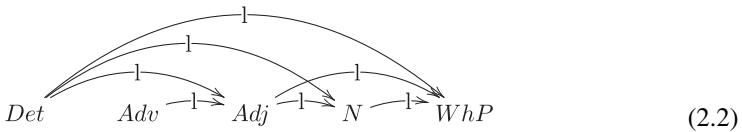
A grammar in PG is a set of all the possible relationships between categories, describing the different constructions. When parsing a given sentence S , assessing a property of the grammar consists in verifying whether the relations between two categories corresponding to words of S are satisfied or not. We present an overview of each type of property.

Linearity: This property implements the same kind of linear precedence relationship as proposed in generalized phase structure grammar [Gazdar et al., 1985]. For example, the nominal construction in English must follow the linearity properties:

$$Det \prec Adj; \quad Det \prec N; \quad Adj \prec N; \quad N \prec WhP; \quad N \prec Prep \quad (2.1)$$

Note that relationships are expressed directly between the lexical categories. As such, the $N \prec Prep$ property indicates precedence between these two categories regardless of their other dependencies. The following example illustrates

the linearity relationships in the nominal construction “*The very old reporter who the senator attacked*”:



In general, properties are also used to control attribute values. For example, one can distinguish linearity properties between the noun and the verb, depending on whether *N* is subject or object by specifying this value in the property itself:

$$N[\text{subj}] \prec V; \quad V \prec N[\text{obj}] \quad (2.3)$$

Co-occurrence: This property typically represents subcategorization, implementing the situation in which two categories must be realized together. An example of co-occurrence within a verbal construction concerns nominal and prepositional complements of ditransitive verbs, which are represented by means of the following properties:

$$V \Rightarrow N; \quad V \Rightarrow \text{Prep} \quad (2.4)$$

It should be noted that co-occurrence not only represents complement-type relationships it can also include co-occurrence properties directly between two categories independent from the head. For example, the indefinite determiner is not generally used with a superlative:

- a. *The most interesting book of the library*
- b. **A most interesting book of the library*

In this case, there is a co-occurrence relation between the determiner and the superlative, which is represented by the property:

$$\text{Sup} \Rightarrow \text{Det}[\text{def}] \quad (2.5)$$

Exclusion: In some cases, restrictions on the possible co-occurrence between categories must be expressed (e.g., lexical selection, concordance). The following properties describe some restrictions in nominal constructions:

$$Pro \otimes N; \quad N[\text{prop}] \otimes N[\text{com}]; \quad N[\text{prop}] \otimes \text{Prep}[\text{inf}] \quad (2.6)$$

These properties stipulate that a pronoun and a noun, a proper noun and a common noun, and a proper noun and an infinitive construction introduced by a preposition cannot be realized simultaneously.

Uniqueness: Certain categories cannot be repeated inside a governing domain. More specifically, categories of this kind cannot be instantiated more than once in a given domain. The following example describes the uniqueness properties for nominal constructions:

$$\text{Uniq} = \{\text{Det}, \text{Rel}, \text{Prep}_{[\text{inf}]}, \text{Adv}\} \quad (2.7)$$

These properties are classical for the determiner and the relative pronoun. They also specify here that it is impossible to duplicate a prepositional construction that introduces an infinitive (“*the will to stop*”) or a determinative adverbial phrase (“*always more evaluation*”).

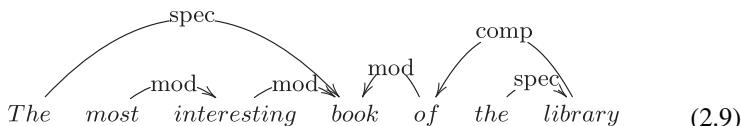
Dependency: This property describes syntax-semantics relationships between categories, indicating that the dependent category complements the governor and contributes to its semantic structure. Dependency relationships are type-based, following a type hierarchy, making it possible to vary the level of precision of the relationship, from the most general to the most specific. These types and subtypes correspond to a classical syntactic relationship:

- dep:** Generic relationship, indicating dependency between a constructed component and its governing component.
- mod:** Modification relationship (typically an adjunct).
- spec:** Specification relationship (typically *Det-N*).
- comp:** The most general relationship between a head and an object (including the subject).
- subj:** Dependency relationship describing the subject.
- obj:** Dependency relationship describing the direct object.
- iobj:** Dependency relationship describing the indirect object.
- xcomp:** Other types of complementation (e.g., between *N* and *Prep*).
- aux:** Relationship between the auxiliary and the verb.
- conj:** Conjunction relationship.

Dependency is noted \rightsquigarrow , possibly completed with the dependency subtype as an index. The following properties indicate the dependency in nominal constructions:

$$\text{Det} \rightsquigarrow_{\text{spec}} N[\text{com}]; \quad \text{Adj} \rightsquigarrow_{\text{mod}} N; \quad \text{WhP} \rightsquigarrow_{\text{mod}} N \quad (2.8)$$

The following example illustrates some dependencies in a nominal construction:



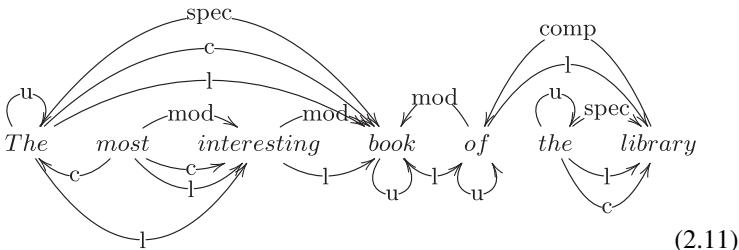
This schema illustrates the specification relationship between the determiner and the noun, the modification relationship between the adjectival and prepositional constructions, and the modification between the adverb and the adjective inside the adjectival construction.

Example: Each property as defined earlier corresponds to a certain type of syntactic information. In PG, the description of syntactic units or linguistic phenomena (chunks, constructions) in the grammar consists of gathering all the relevant properties into a set. Schema 2.10 summarizes the properties describing the nominal construction.

$Det \prec \{Det, Adj, WhP, Prep, N\}$	$Det \rightsquigarrow_{spec} N$
$N \prec \{Prep, WhP\}$	$Adj \rightsquigarrow_{mod} N$
$Det \Rightarrow N[com]$	$WhP \rightsquigarrow_{mod} N$
$\{Adj, WhP, Prep\} \Rightarrow N$	$Prep \rightsquigarrow_{mod} N$
$Uniq = \{Pro, Det, N, WhP, Prep\}$	$Pro \otimes \{Det, Adj, WhP, Prep, N\}$
	$N[prop] \otimes Det$

(2.10)

A syntactic description, instead of being organized around a specific structure – a tree, for example – consists of a set of independent properties together with their status (satisfied or violated). The graph in Schema 2.11 illustrates the PG description of the nominal construction: *The most interesting book of the library*, where l represents linearity, u is uniqueness, and c is co-occurrence.



In PG, a syntactic description is therefore the graph containing all the properties of the grammar that can be evaluated for the sentence to be parsed. As illustrated in the example, this property graph represents explicitly all the syntactic characteristics associated with the input, and each is represented independent from the others.

2.4 Chunks, Constructions, and Properties

As we noted in Section 2.2 An Interdisciplinary View of Language Processing, many observations tend to show how important chunks and constructions can

be in LP architecture. Our hypothesis is that two different types of processing coexist: one serial, incremental, word-by-word and compositional (the classical LP organization in the literature) and another shallow, based on chunks or constructions, recognized as a whole, giving, when possible, direct and global access to the meaning. This hypothesis is supported by several observations, showing the existence of such units in particular when studying the brain correlates of language processing. Moreover, some basic properties (typically agreement) are identified at a very early stage, indicating an automatic and low-level process. We will first explain what these basic properties are and then how chunks or constructions can be recognized on the basis of the properties they contain.

2.4.1 Basic Properties

The different properties presented in earlier sections can be assessed directly when processing a sentence: for each set of categories, it is possible to verify whether some properties link them in the grammar and whether, in the specific context of their realization in the sentence, they are satisfied or not. A property plays exactly the role of a constraint, describing an input consists in assessing the properties, assigning them a truth value.

Two types of properties can be distinguished, according to the way they can be evaluated and their sensitivity to the context [Blache and Rauzy, 2004]. More precisely, a property can be assessed as soon as the categories they concern are recognized in a sentence. The difference between the two types of properties is that in one case, their satisfaction remains the same whatever the window of words taken into consideration, and in the other case, this value can vary depending on the window (being sensitive to the context).

- *Success-monotonic properties:* When a property between two categories becomes satisfied, it remains satisfied for the entire sentence. For example, the linearity between *most* and *interesting* in Schema 2.11 holds as soon as it can be assessed, and remains satisfied until the end, whatever the span of words.

More formally, the linearity relationship $a \prec b$ is satisfied in the sequence of words $s = [\gamma, a, b, \eta]$, whatever the composition of γ and η . Two types of properties are success-monotonic: *linearity* and *co-occurrence*.

- *Success-nonmonotonic properties:* A property can be satisfied locally and become violated at a larger span: the evaluation of a property depends on the set of categories taken into account. For example, an *exclusion* relationship between the words a and d is satisfied within the set of words $s1 = \{a, b, c\}$, but false when adding a new category d to this sequence $s2 = \{a, b, c, d\}$. In

this case, it is necessary to specify the sequence (or the partition) for which the constraint is evaluated.

Success-monotonic properties are computationally simpler than nonmonotonic properties because they do not need to be re-evaluated at each step; when such a property becomes satisfied, this assessment cannot be reconsidered. We characterize these types of properties as basic. They are low-level properties, automatically assessed at an early stage in the brain. Moreover, they encode the two types of information used when evaluating transition probabilities between categories (linearity and co-occurrence), reinforcing the proposal to consider them at a first level.

2.4.2 Chunks from Properties

We have seen how to recognize and assess properties. The question is now how they can be used to identify a chunk or a construction. Several experiments have shown that some syntactic properties can be assessed very early, without any deep and precise analysis (see in particular Pulvermüller et al., 2008). In our hypothesis, they correspond to “basic properties,” which can be evaluated based on the immediate context. The next challenge is to learn how to identify higher-level organizations such as chunks and constructions.

Our proposal relies on the idea that in some cases, there exists a link between properties: the existence of some properties can also activate other properties. For example, the verification of a linearity property between a *Det* and a *N* activates a dependency relationship between them. This same kind of relationship exists in lexical selection, collocations, etc.: the realization of a given word activates or predicts that of another one.

As a consequence, the description of a construction in the grammar consists in two types of information: the set of properties and the identification of those basic properties that can activate other ones. We propose to add this information to the representation of the properties by means of a new argument encoding the properties that can be linked to the current as follows:

```
<id, type, source, target, weight, linked_props>
```

The *linked_props* argument is a set of indexes that point to other properties describing the same construction. For example, the dependency relationship between a preposition and a noun depends on the linearity: if *Prep* \prec *N*, then *Prep* is the head and *N* depends on it. Reciprocally, when *N* \prec *Prep*, the *Prep* depends on *N*. These relationships between properties are represented as follows:

```
<1, lin, Prep, N, H, {}> <2, comp, N, Prep, S, {1}>
<3, lin, N, Prep, H, {}> <4, mod, Prep, N, S, {3}>
```

The example of the ditransitive construction can be implemented in the same manner, specifying different dependency types¹ according to the form (the first noun is the indirect object, the second the direct):

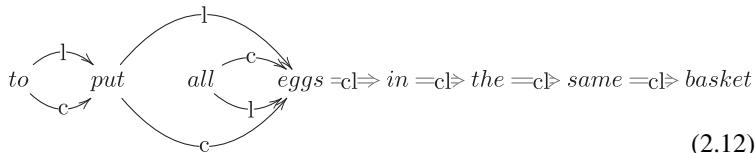
```
<1, lin, V[dit], N1, H, {}> <4, iobj, N1, V, H, {1,2,3}>
<2, lin, V[dit], N2, H, {}> <5, obj, N2, V, H, {1,2,3}>
<3, lin, N1, N2, H, {}>
```

2.4.3 Processing Idioms

The activation mechanism based on linked properties can be generalized to the processing of other types of constructions. We know, for example, that processing idioms (see Vespignani et al., 2010 and Rommers et al., 2013), is done in two different steps, before and after the word starting from which the idiom is recognized (called the *recognition point*, or RP). Before the RP, the processing consists of assessing basic properties. At the RP, the idiom is recognized, its meaning is globally accessed, without any need to analyze the rest of the idiom. All the remaining words become fully predictable. In terms of properties, this means that a set of mandatory co-occurrences as well as linearity between the list of words is activated.

This phenomenon can be implemented with the mechanism of *linked properties*: reaching the RP means having already assessed a certain amount of basic properties, relating the initial words of the idiom. Recognizing the RP consists of inferring a set of linked properties from the basic ones.

The following figure illustrates this mechanism for the idiomatic expression *to put all eggs in the same basket*:



In this idiom, the RP is at the word *eggs*. Before the RP, the basic properties are assessed, linking the first words of the idioms. After this point, all the other properties can be automatically inferred, as well as the association of a global meaning. The rest of the process consists only of verifying whether the prediction matches the remaining words.

The property-based description of this idiom can be implemented with the following properties:

¹ At this stage, to be as generic as possible, representation of the dependencies usually requires underspecification.

- (1) $\text{put} \prec \text{all}$
- (2) $\text{all} \prec \text{eggs}$
- (3) $\text{put} \Rightarrow \{\text{in}, \text{one}, \text{basket}\}$
- (4) $\text{eggs} \prec \text{in} \prec \text{one} \prec \text{basket}$
- (5) $\text{sem}(\text{put}) = [[\text{risk_losing_everything}]]$

In this description, we only describe the basic linearity and co-occurrence properties. The RP is implemented by factorized properties (3) and (4). The general mechanism is described by the following formula, indicating that properties (3), (4), and (5) can be inferred directly from the basic properties (1) and (2):

$$(1 \wedge 2) \Rightarrow (3 \wedge 4 \wedge 5)$$

Note that the semantics of the idiom is represented by a denotation attached, arbitrarily, to the verb (the idiomatic construction being verbal in this case).

2.5 The Hybrid Architecture

The language-processing architecture we propose is an alternative to the classical incremental, modular, and serial organization. We think that, instead of processing word-by-word by trying to integrate each new word into a partial structure and interpreting the result compositionally (the meaning of the whole being a function of each component), it is preferable to propose a flexible architecture, more in line with what is observed in human LP.

2.5.1 General Organization

The first general idea is that processing is not strictly incremental and meaning access not compositional. We have described a basic processing level in which words are gathered into larger units. There are some situations, typically constructions, in which meaning is assessed directly. This means that two different types of processing are juxtaposed: the first type, which relies on low-level mechanisms and is considered the default level, directly identifies chunks that offer the potential for global processing. The second type is classical, word-by-word, serial, and compositional, and is applied when the first is not possible.

Generally speaking, we consider that the first level of processing is superficial and delay as much as possible the interpretation. Whenever possible, larger units grouping several words are built. Such groups make it possible to gather different sources of information, preparing a first level of interpretation. In some cases, they even constitute entire constructions, offering the possibility to directly access the meaning. This first level of processing is supposed to be done automatically, on the basis of low-level mechanisms.

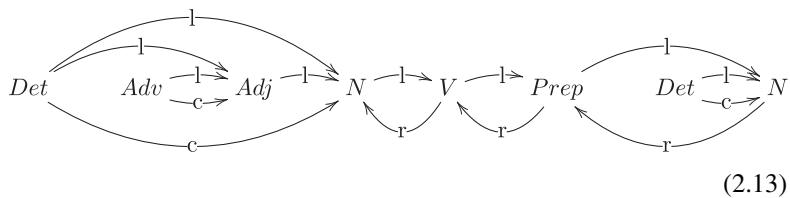
These units are stored in the working memory, together with their interpretation when it exists. Global interpretation then groups the different local meanings together. When no grouping is possible, then words (instead of groups) are stored in the memory. In both cases, the interpretation process is only done after gathering a certain amount of information (or when reaching the maximal capacity of the working memory).

As a result, different types of objects coexist: words, chunks, and constructions. The existence of units grouping words facilitates the processing: these objects are recognized by means of low-level mechanisms and offer the possibility to directly contribute to the meaning. We propose a method for identifying these units.

2.5.2 Recognizing chunks

Chunks are set of words, usually adjacent (but not necessarily), linked by tight morphosyntactic relationships (typically *Det-N*). As noted earlier, such relationships mainly correspond to what we call basic properties, that is, linearity and co-occurrence. When parsing an input, processing a new word consists of checking such properties with adjacent words. The resulting graph makes it possible to identify subgroups, formed by the set of words that are connected by such properties. When looking at a constraint graph obtained from basic properties, such subgroups can be immediately identified: they correspond to the complete subgraphs (the set of nodes in a graph that are directly connected).

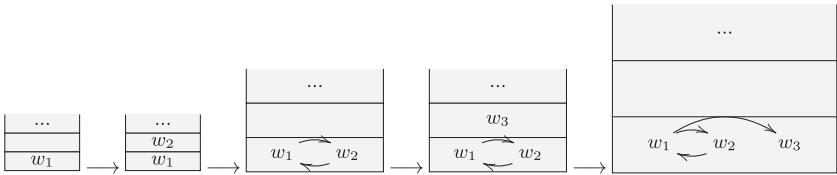
The example in Schema 2.13 illustrates a constraint graph, containing basic properties only:



In this graph, we can identify several subset of nodes that are all connected. For example, the subset *Adv-Adj* forms a complete subgraph, but not *Det-Adv-Adj*: in this last case, *Det* is not connected with *Adv*. A complete subgraph is then made of words with morphosyntactic relationships that corresponds to our definition of chunks. The list of complete subgraphs in Schema 2.13 is: *Adv-Adj*; *Det-Adv-Adj-N*; *N-V*; *V-Prep*; *Prep-N*; *Det-N*. This corresponds to the list of the chunks, identified only by means of basic properties.

Chunk recognition constitutes the first level of processing. In the following, we leave aside the question of word segmentation and recognition during human LP (even though it remains an open question) and consider the initial

input as a set of words. The processing consists of scanning the input word by word. The initial word of the input is simply stored in the working memory, which is made of several buffers (represented here as a stack). The next step consists of scanning a new word, pushing it into the stack of buffers, and then determining whether it can be connected by basic properties with the word in the lower buffer. When the new word can be linked to the previous word, the two words form a chunk, which is pushed into the buffer. In the same way, a word can be linked to an existing chunk, forming a new chunk to replace the previous one, as illustrated in Figure 2.8.



2.5.3 Recognizing Constructions

The identification of constructions can be explained thanks to linked properties and cohesion evaluation. They constitute a distinct mechanism that makes it possible, starting from a chunk, to identify the type of the construction and complete its description with new properties.

A chunk is made of words connected with basic properties. In some cases, this set of properties can be associated to linked properties, resulting in the inference of new relations completing the constraint graph.

As shown in Figure 2.8, this mechanism explains the effect of the recognition point in idiom processing. Before this point, the preceding words are processed as explained in Section 2.4.3. Processing Idioms, building chunks when possible. When reaching the word corresponding to the RP, a set of linked properties is directly assessed. We obtain then a new constraint graph which also bears for idioms a complete interpretation; in this case, the entire set of properties describing this specific construction is formed by linked properties, making it possible to infer directly the description and its interpretation.

For other types of constructions, with a certain level of flexibility [Goldberg, 2006], only subparts of the properties are linked and can be automatically inferred from the basic ones. In this situation, an evaluation of the graph density completes the mechanism. The constraint graph, completed by the possible linked properties, is analyzed. If its density value reaches a certain threshold, then the construction is recognized and the entire set of linked properties is activated. This mechanism gives direct access to the meaning associated with the construction, to be completed when scanning the rest of the input.

2.5.4 Light Parsing with Chunks and Constructions

Recognizing chunks and a fortiori constructions facilitates language processing because it allows direct access to a certain interpretation thanks to basic properties. In our approach, this constitutes the first level of processing based on the assessment of basic properties that can trigger the inference of other types of properties thanks to the mechanism of linked properties. In this hypothesis, constructions are encoded in the memory as *recurrent networks*, encoding directly linked properties. This view fits with the *MUC model* (memory, unification, control) proposed by Hagoort [Hagoort, 2005; Hagoort, 2013] in which memory contains lexical building blocks that encode complex lexical entries, including syntactic and semantic relations. MUC is in line with several linguistic theories such as head-driven phrase structure grammars [Pollard and Sag, 1994] or tree-adjoining grammars [Joshi and Schabes, 1997] in which most syntactic and semantic information is encoded in the lexicon. The memory stores such units when the unification component is in charge of integrating them. In our model, the linked properties are stored in the memory together with lexical units. The unification component can then directly assemble units thanks to a simple mechanism: basic properties assessment plus linked properties inference.

Finally, light processing architecture distinguishes two levels of unification during sentence processing:

- *Light level*: Used as default, storing words in the working memory, assembling them into chunks, inferring linked properties and activating constructions when possible.
- *Deep level*: Used when the light level does not lead to interpretation. The processing is classical: strictly incremental and serial, interpretation being built compositionally, starting from a syntactic structure.

2.6 Conclusion

Sentence processing is fast and happens in real time despite the complexity of linguistic mechanisms. One explanation is that language makes use of frequent structures or patterns that can be learned. This justifies the use of probabilistic approaches that is at work today in most LP models. However, several experiments have shown that even some types of rare structures can be processed in real time [Pulvermüller, 2010]. We have described in this chapter a new way of representing linguistic information, based on properties, and we have shown how two types of such properties can be distinguished, according to the way they can be verified. Some properties, known as basic properties, are assessable simply and directly. Moreover, we have shown how properties, whatever their type, can be linked and directly inferred from each other. This

mechanism (basic assessment+inference) is the basis for recognizing chunks and constructions. It constitutes the first level of parsing in our model, based on *light parsing*. This processing mechanism is the default one, explaining why language processing can be often shallow but fast: interpretation can be directly accessed thanks to such basic mechanism. In some difficult and complex cases, light parsing does not lead to any interpretation. In such cases, a classical *deep parsing*, incremental and serial, is used.

The *light and deep parsing model*, therefore, constitutes a candidate for a new language processing architecture, explaining why human LP is efficient and opening the way to new types of experiments in neurolinguistics.

References

- Abney, S. (1991). Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics*, Dordrecht; Boston: Kluwer Academic Publishers, pages 257–278.
- Adriaens, G. and Hahn, U., editors (1994). *Parallel Natural Language Processing*. Norwood, NJ: Ablex Publishing Corporation.
- Altmann, G. T. M. and Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4):583–609.
- Baldwin, T., Dras, M., Hockenmaier, J., King, T. H., and van Noord, G. (2002). The impact of deep linguistic processing on parsing technology. In *Proceedings of IWPT-2007*.
- Balfourier, J.-M., Blache, P., and Rullen, T. V. (2002). From shallow to deep parsing using constraint satisfaction. In *Proc. of the 6th Int'l Conference on Computational Linguistics (COLING 2002)*.
- Blache, P. (2000). Property grammars and the problem of constraint satisfaction. In *Linguistic Theory and Grammar Implementation*, ESSLLI 2000 workshop.
- Blache, P. and Rauzy, S. (2004). Une plateforme de communication alternative. In *Actes des Entretiens Annuels de l'Institut Garches*, pages 82–93, Issy-Les-Moulineaux, France.
- Cappelle, B., Shtyrov, Y., and Pulvermüller, F. (2010). Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain and Language*, 115(3), 189–201.
- Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht; Cinnaminson, NJ: Foris Publications.
- Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2001). Minimal recursion semantics: An introduction. In *Language and Computation (L&C)*, volume 1, pages pp. 1–47. Oxford: Hermes Science Publishing.
- de Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford Parser v. 3.5.2.
- Ferreira, F. and Patson, N. D. (2007). The “good enough” approach to language comprehension. *Language and Linguistics Compass*, 1(1).
- Fillmore, C. J. (1988). The mechanisms of “construction grammar.” In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.
- Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

- Fodor, J. D. and Ferreira, F. (1998). *Reanalysis in Sentence Processing*, Dordrecht; Boston: Kluwer Academic Publishers.
- Fodor, J. and Inoue, A. (1998). Attach anyway. In Fodor, J. and Ferreira, F., editors, *Reanalysis in Sentence Processing*. Dordrecht; Boston: Kluwer Academic Publishers.
- Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(22):78–84.
- Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, 91(4):1357–1392.
- Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985). *Generalized Phrase Structure Grammar*. Oxford: Blackwell.
- Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistic Complexity. In Marantz, A., Miyashita, Y., and O’Neil, W., editors, *Image, Language, Brain*, pages 95–126. Cambridge, MA: MIT Press.
- Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
- Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.
- Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford, UK: Oxford University Press.
- Gorrell, P. (1995). *Syntax and Parsing*. Cambridge, UK: Cambridge University Press.
- Grodner, D. J. and Gibson, E. A. F. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29:261–291.
- Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, 9(9):416–423.
- Hagoort, P. (2013). Muc (memory, unification, control) and beyond. *Frontiers in Psychology*, 4:416.
- Jackendoff, R. (2007). A parallel architecture perspective on language processing. *Brain Research*, 1146(2–22).
- Jindal, P., Roth, D., and Kale, L. (2013). Efficient development of parallel nlp applications. Technical report, Tech. Report of IDEALS (Illinois Digital Environment for Access to Learning and Scholarship).
- Joshi, A. K. and Schabes, Y. (1997). Tree-adjoining grammars. In Rozenberg, G. and Salomaa, A., editors, *Handbook of Formal Languages, volume 3: Beyond Words*, pages 69–124. New York: Springer.
- Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, 1(6).
- Keller, F. (2010). Cognitively plausible models of human language processing. *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67.
- Kutas, M., Petten, C. K. V., and Kluender, R. (2006). Psycholinguistics electrified ii: 1994–2005. In Gernsbacher, M. A. and Traxler, M., editors, *Handbook of Psycholinguistics*, pages 659–724. Boston: Elsevier.
- Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. Boston: MIT Press.
- MacDonald, M., Pearlmutter, N., and Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101:676 –703.

- Marslen-Wilson, W. and Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition*, 8:1–71.
- Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammars*. Center for the Study of Language and Information Publication (CSLI), Chicago: Chicago University Press.
- Pulvermüller, F. (2010). Brain embodiment of syntax and grammar: Discrete combinatorial mechanisms spelt out in neuronal circuits. *Brain and Language*, 112(3):167–179.
- Pulvermüller, F., Shtyrov, Y., Hasting, A. S., and Carlyon, R. P. (2008). Syntax as a reflex: Neurophysiological evidence for early automaticity of grammatical processing. *Brain and Language*, 104(3):244–253.
- Rauzy, S. and Blache, P. (2012). Robustness and processing difficulty models. A pilot study for eye-tracking data on the French treebank. In *Proceedings of the 1st Eye-Tracking and NLP workshop*.
- Rayner, K. and Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, 80(1):4–9.
- Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: Evidence from idiom comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776.
- Schuett, S., Heywood, C. A., Kentridge, R. W., and Zihl, J. (2008). The significance of visual information processing in reading: Insights from hemianopic dyslexia. *Neuropsychologia*, 46(10):2445–2462.
- Spivey, M. J. and Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 24:1521–1543.
- Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689.
- Sturt, P. and Lombardo, V. (2005). Processing coordinated structures: Incrementality and connectedness. *Cognitive Science*, 29(2).
- Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, 36(1):201–216.
- Uszkoreit, H. (2002). New chances for deep linguistic processing. In *proceedings of COLING-02*.
- Vespignani, F., Canal, P., Molinaro, N., Fonda, S., and Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.
- Werning, M., Hinzen, W., and Machery, E. (2012). *The Oxford Handbook of Compositionality*. Oxford, UK: Oxford University Press.

3 Decoding Language from the Brain

Brian Murphy, Leila Wehbe, and Alona Fyshe

Abstract

In this paper we review recent computational approaches to the study of language with neuroimaging data. Recordings of brain activity have long played a central role in furthering our understanding of how human language works, with researchers usually choosing to focus tightly on one aspect of the language system. This choice is driven both by the complexity of that system, and by the noise and complexity in neuroimaging data itself. State-of-the-art computational methods can help in two respects: in teasing more information from recordings of brain activity and by allowing us to test broader and more articulated theories and detailed representations of language tasks. In this chapter, we first set the scene with a succinct review of neuroimaging techniques and what they have taught us about language processing in the brain. We then describe how recent work has used machine learning methods with brain data and computational models of language to investigate how words and phrases are processed. We finish by introducing emerging naturalistic paradigms that combine authentic language tasks (e.g., reading or listening to a story) with rich models of lexical, sentential, and suprasentential representations to enable an all-round view of language processing.

3.1 Introduction

The study of language, like other cognitive sciences, requires of us to indulge in a kind of mind reading. We use a variety of methods in an attempt to access the hidden representations and processes that allow humans to converse. In formal linguistics intuitive judgments by the theorist are used as primary evidence – an approach that brings well-understood dangers of bias (Gibson and Fedorenko, 2010), but in practice can work well (Sprouse et al., 2013). Aggregating judgments over groups of informants is widely used in cognitive and computational

linguistics, through both experts in controlled environments and crowdsourcing of naive annotators (Snow et al., 2008). Experimental psycholinguists have used a range of methods that do not rely on intuition, judgments, or subjective reflection, such as the speed of self-paced reading, or the order and timing of gaze events as recorded with eye-tracking technologies (Rayner, 1998).

Brain-recording technologies offer a different kind of evidence, as they are the closest we can get empirically to the object of interest: human cognition. Despite the technical challenges involved, especially the complexity of the recorded signals and the extraneous noise that they contain, brain imaging has a decades-long history in psycholinguistics. Particular patterns of electrophysiological activity are associated with processing difficulties in the meaning and structure of sentences, and relative changes in blood flow can reveal parts of the brain whose activity is modulated by the complexity of a language processing task (see Section 3.2). These experimental approaches usually frame theoretical questions in terms of a small number of categories (e.g., familiar words versus obscure words to look at the processing associated with lexical retrieval), and in terms of tasks that try to “stretch” or “break” language to see how it functions (e.g., through the use of ill-formed sentences).

In this chapter, we present an additional stream of recent work that uses computational modeling of both language and brain activity to build on this research.

In Section 3.3, we describe studies that explore the breadth and depth of the human lexicon. Models from computational lexicography and the word vector space/embedding literature are employed to empirically model the various dimensions of meaning along which words can differ, which are common to many early (Katz and Fodor, 1963) and current theories (Barsalou and Wiemer-Hastings, 2005; Baroni and Lenci, 2010) of word meaning. By employing distributional semantic theory as a “prior”, we can use computational models to separate those aspects of brain activity that are related to word meaning from those related to other aspects of the experimental task or extraneous noise.

In Section 3.4, we look above the level of the word, at how lexical units combine, in real time, to form short phrases. As in the work with single words, we can take advantage of existing theories of language to characterize the representations produced by compositional processes. This takes advantage of the fact that huge quantities of textual data are cheaply available to build and evaluate fine-grained models, and then these models can be tested against the expensive and limited collections of brain data. Again, we differentiate the brain activation attributed to the act of composition from the brain activity attributable to the composed semantics of a phrase.

Most recently, there has been a movement in cognitive neuroscience toward the use of naturalistic tasks, which are claimed to be more ecologically valid

(Willems, 2015). In Section 3.5, we describe some experiments that use natural reading and listening tasks to elicit holistic and realistic language processing, without resorting to constructed stimuli with hand-engineered syntactic or semantic errors. A combination of tools from computational linguistics and crowd annotation allows us to build detailed multilevel models of the perceptual, structural, and semantic work involved in understanding a real narrative. As with the word- and phrase-level work, the use of such a model brings several advantages. Computational modeling of the relationship between word features and brain areas differentiates the brain activity driven by language processing from irrelevant brain activity. Closer inspection of those sensitivities can tell us which brain areas and which parts of the time course are tied to particular subprocesses. And the generative nature of the models allows us to perform the “mind reading” trick of estimating (imperfectly, but at a level clearly above chance) what word, phrase, or sentence a person is processing at a given moment.

In this review we concentrate on research that uses recordings of brain activity and computational modeling as a tool to probe how language functions in the mind, rather than work that uses language as a probe to understand brain function. There is also a very large body of work that develops models of brain activity that build in it spatial, temporal, and network characteristics; that work is not covered here.

3.2 Grounding Language Architecture in the Brain

The earliest investigations of language in the brain began in the 1800s, when Paul Broca and Karl Wernicke studied patients with brain injuries that affected their ability to communicate (Bear et al., 2007). Broca’s patients exhibited partial or complete loss of language abilities (aphasia), and their pattern of brain damage prompted him to conclude that language is controlled by a single hemisphere of a person’s brain, almost always the left hemisphere (Dronkers et al., 2007).¹ Broca’s work also led him to identify a region of the brain in the posterior inferior left frontal gyrus (“Broca’s area,” see Figure 3.1) associated with a particular variety of aphasia. Nonfluent aphasia, also called Broca’s aphasia or expressive aphasia, is characterized by the impairment of language production, while comprehension and general cognition remain intact.

Wernicke found that lesions to a different left hemisphere region (posterior superior temporal gyrus, or “Wernicke’s area,” see Figure 3.1) led to a distinct

¹ Throughout this chapter, we refer to the typical left-lateralized localization of language areas in the brain of a right-handed person. Left-handed people (and, indeed, some right-handed people) have language centers located in the right hemisphere of the brain.

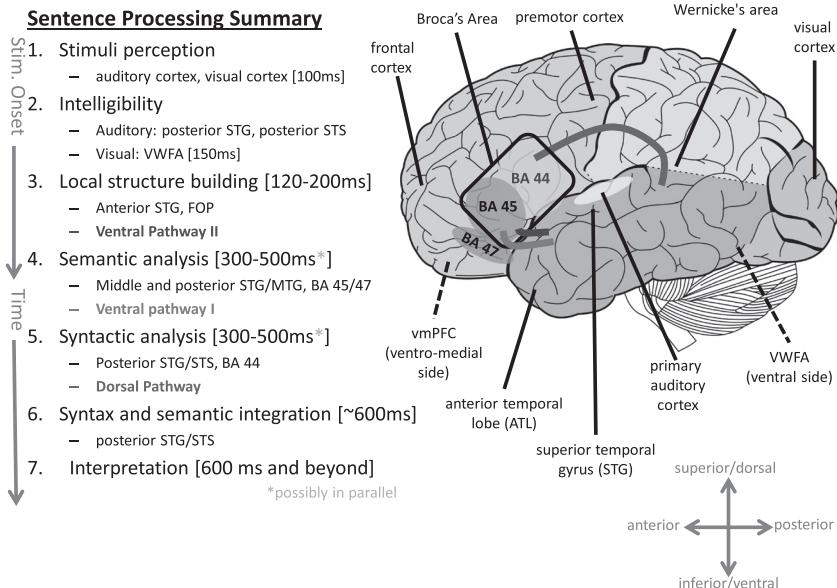


Figure 3.1 An overview of the location and timing of language processing (for sentences) in the brain. Posterior refers to areas toward the back of the head; anterior, toward the front; superior/dorsal, toward the top of the head; inferior/ventral, toward the bottom. Medial refers to locations toward the middle of the brain, where the two hemispheres meet. A gyrus refers to a ridge of the cortex, and a sulcus is the fold between gyri. BA, Brodmann area; STG, superior temporal gyrus; STS, superior temporal sulcus; VWFA, visual word form area; FOP, frontal operculum (just ventral to Broca's area); MTG, middle temporal gyrus (between superior and inferior gyri); vmPFC, ventromedial prefrontal cortex. Adapted from Friederici (2011).

pattern of language impairment. Fluent aphasia (also jargon aphasia or Wernicke's aphasia) is characterized by the easy production of language that is mostly nonsensical or wandering. Intonation and speed of speech are usually normal, and if one ignores the content of the utterance, the speech can seem quite typical. Patients often have difficulty following verbal instructions, indicating that language understanding is also affected. These symptoms have led to the theory that these areas are instrumental in mapping language sounds or written words to semantic content.

Since the 1970s the study of language in the brain has been transformed by brain-imaging technologies, most commonly electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI).

EEG is the oldest of the three brain-imaging technologies discussed here. EEG measures the voltage fluctuations on the scalp, induced by the coordinated firing of groups of brain cells, colloquially referred to as brain waves. The propagation of these postsynaptic electrical fields to the scalp is instantaneous, but is distorted spatially and temporally through low-pass filtering, as it passes through varying densities of tissue and the skull. Scalp signals can be resolved back to underlying brain sources after making some physical assumptions, but even in ideal settings spatial resolution is poor, on the order of 7 mm (Im et al., 2007). On the other hand, EEG gives excellent temporal resolution, usually recorded at hundreds or thousands of samples per second (Hz), whereas the maximum firing rate of neurons is typically about 50 Hz. EEG has an additional advantage among the modalities discussed here in that it requires quite simple equipment (essentially sophisticated signal amplifiers) and so can be used in a range of lab and nonlab environments. Miniaturization of electronics has recently made wearable EEG a reality for both research and consumer uses.

MEG measures the minuscule magnetic field corresponding to the electrical fields detected by EEG. Like EEG, MEG measures the postsynaptic currents in apical dendrites (i.e., the arrival of a new electrical message to a neuron), particularly of cells in the sulci (“valleys” of the cortical folding) (Hansen et al., 2010). The propagation of these magnetic fields is not distorted by passage through the head, and so the signals, although similar in kind to EEG, are cleaner and contain more high-frequency content (a sample MEG recording appears in Figure 3.2). MEG signals can be resolved to locations in the brain with a much higher spatial resolution, on the order of 2–3 mm in ideal conditions (Hamalainen et al., 1993). However MEG is an expensive and complex technology, requiring a magnetically shielded room and supercooling to support superconducting magnetometers.

The imaging technique with the greatest spatial resolution is fMRI, which can achieve resolution as fine as 1 mm. A sample fMRI image appears in Figure 3.3. fMRI measures changes in blood oxygenation in response to increased neuronal activity, called the blood oxygen level-dependent (BOLD) response. Because fMRI depends on the transport of oxygen via blood to the brain, its time constant is governed by the rate at which blood can replenish oxygen in the brain. Although fMRI can acquire images at the rate of about one image per second, the BOLD response can take more than 5 s to reach its peak after a stimulus is shown. Thus, among the three modalities discussed here, fMRI has the worst time resolution and the best spatial resolution.

Neurolinguistic studies generally use carefully controlled comprehension tasks, such as rapid serial visual presentation (presenting a phrase or sentence, word by word at a fixed rate on a screen) or sentence reading followed by questions to ensure that the participant is attending to and processing the

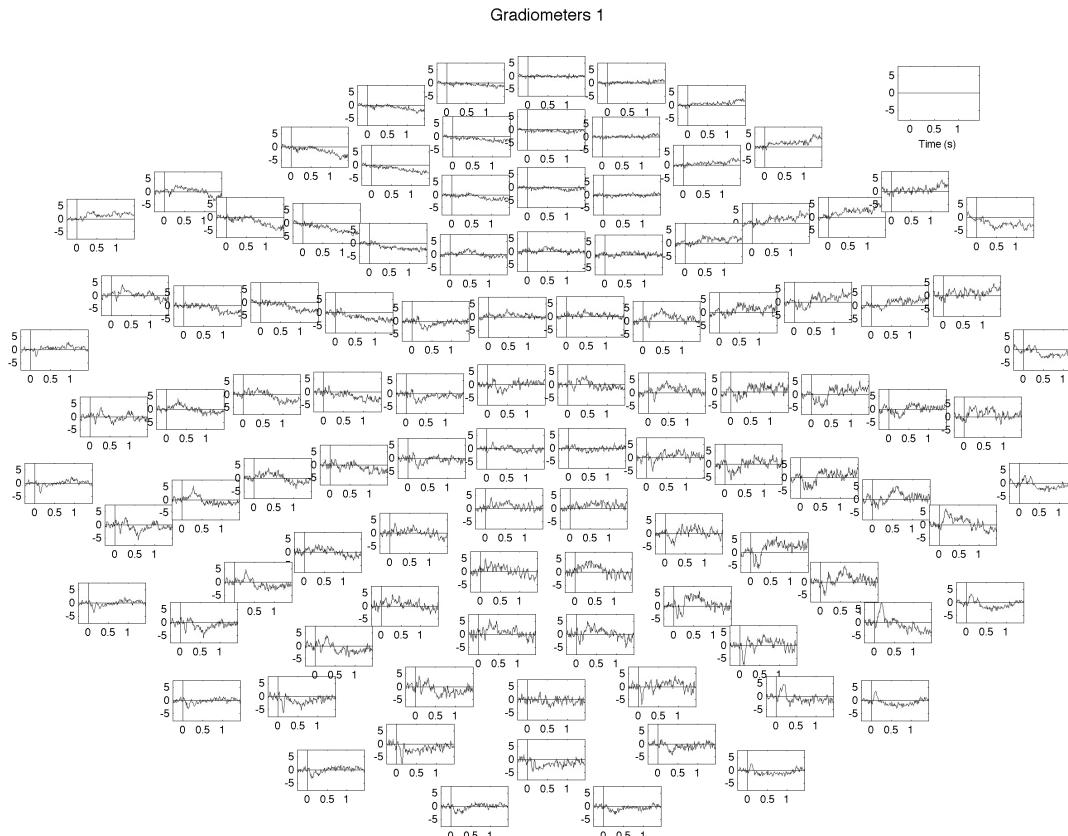


Figure 3.2 An example MEG recording averaged over twenty repetitions of a person reading the word *bear*. Each subplot represents the recording from the first gradiometer at one of the one hundred and two sensor positions on the MEG helmet. For simplicity, the other two hundred and four sensor recordings are not shown. In this diagram, the helmet is oriented as if we are looking down on it from above. The nose of the subject points to the top of the figure, and the left side of figure corresponds to the left hand side of the subject. Time is along the x-axis of each plot and the y-axis corresponds to the gradient of the magnetic field in 10^{-3} T/cm (Fyshe, 2015).

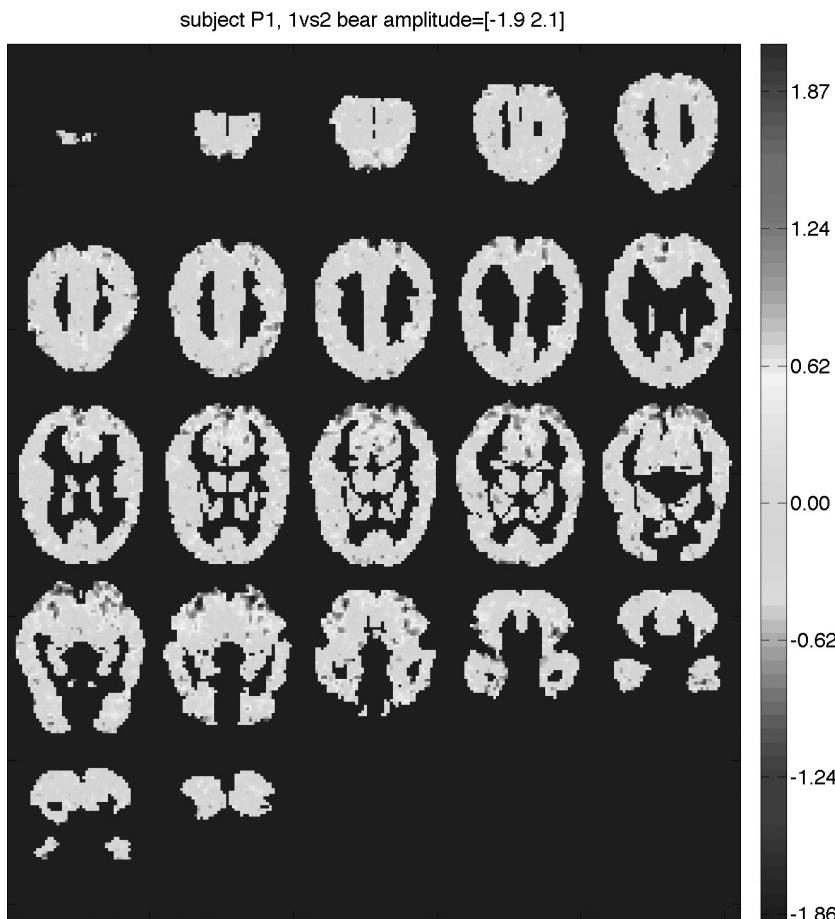


Figure 3.3 An fMRI image averaged over six repetition of a person reading the word *bear*. An fMRI image is three-dimensional, here shown in slices progressing from the top of the head (*top left*) to the bottom of the head (*bottom right*). In each slice, the front of the head points towards the bottom of the figure, and the right side of the subject is shown on the left side of the each image (as if we are viewing the brain of a subject laying face down, from the top of their head). The color of each voxel (pixel in brain space) represents the percent change over baseline of the BOLD response in that brain area (Fyshe, 2015).

materials. This limits the conditions to be analyzed, and avoids machine artifacts due to participant movement. Reading tasks are predominant, as they simplify the question of when (visual) word processing begins. With auditory presentation (Hagoort, 2008; Brennan and Pylykänen, 2012) the articulation of a word can stretch over many hundreds of milliseconds and it is not self-evident

at which point during the word processing can begin, given anticipatory effects (Marslen-Wilson and Tyler, 1980; DeLong et al., 2005).

Neuroimaging work on language started with EEG. Kutas and Hillyard (1980) noted a more negative current in centroparietal sensors triggered by a semantically mismatched ending to a sentence (e.g., *He spread the warm bread with socks*), which became apparent from about 400 ms after the critical incongruous word appeared. This N400 was originally thought to be a reaction to semantically incompatible words, but it has since been shown that the N400 can be evoked by sentences with semantically less predictable words. For example, *Jenny put the sweet in her (mouth/pocket) after the lesson* elicits an N400 for the word *pocket*. Although *pocket* is a semantically sound word choice, it is judged less probable than the alternative (*mouth*). It has also been shown that the N400 can appear before the incongruent word if the indefinite article (*a/an*) does not match the predicted noun. For example, an N400 will occur for the word *an* in the sentence *It was breezy, so the boy went to fly an kite* because the word *kite* is so strongly predicted and *an* is the wrong indefinite article (DeLong et al., 2005). The N400 can also be evoked in the context of arbitrary lists of words, where the magnitude of the effect is larger for infrequent words, and smaller for words that have been semantically primed by previous words presented. This effect has led to the interpretation that the N400 indexes the cognitive load involved in retrieving a word from the mental lexicon (Kutas and Federmeier, 2011).

In contrast to the N400, the P600 is characterized by a positive-going current that peaks around 600 ms after stimulus onset, also in centroparietal sensors. Typically, the P600 is seen at a sentence position where there is a syntactic violation (e.g., word order mistakes, plural verb disagreement, grammatical gender mismatch) (Kuperberg, 2007). However, under certain circumstances a P600 can be evoked even when the syntax of a sentence is correct. For example, the sentence *Every morning the eggs would eat toast for breakfast* will induce a P600 for the underlined word “eat,” although the sentence is syntactically sound, and elicits no N400, although the sentence is semantically incongruent. This phenomenon was called a “semantic illusion” because it was seen as “fooling” the reader into thinking that the word is semantically sound due to a strong conceptual link (Hoeks et al., 2004).

Because of their excellent time resolution, EEG and MEG have provided insights into the manner and order of processing in language comprehension. One major debate in linguistics is on the degree of modularity and serialization in language processing. For example, early models by Friederici (2002) posited a strictly serial model consistent with syntax-central and modular models of language (Fodor, 1983; Chomsky, 1995). The early timing of semantic effects evidenced in the N400 and the fact that it can appear also in nonsentence final positions suggest that syntax analysis does not always strictly precede

semantics, and later studies from the Hagoort lab demonstrated that both encyclopedic knowledge and discourse-specific facts are integrated online into computations of semantic correctness (Hagoort et al., 2004; Özyürek et al., 2007).

The fine spatial resolution of fMRI and MEG have allowed us to study specific regions of the brain during language processing. Levels of processing can be disentangled by comparing the brain activity elicited by different kinds of language materials. For instance, brain areas that are active for real words, but not for pseudowords, might be assumed to be involved in lexical retrieval (McCandliss et al., 2003; Salmelin, 2007). Those areas active for both but not active for a nonlinguistic symbols might be assumed to be involved in reading. And those areas active for phrases and sentences but not for single words presumably play a role in syntactic or semantic composition.

Following this analytical paradigm, large swathes of cortex are implicated as specialized for language processing (Fedorenko et al., 2012). The temporal lobe is broadly involved (including Wernicke's area), is almost always reported as having left-hemisphere predominance, and is thought by many to be primarily responsible for lexical processing. The inferior frontal gyrus (which includes Broca's area) and neighboring left-hemisphere areas are thought to be involved in processing sentence structure and meaning, although this is an active area of research (Hagoort, 2005; Friederici, 2011).

Models of language processing have taken inspiration from vision research (Hickok and Poeppel, 2004, 2007; Friederici, 2011), which has broken visual processing into two streams: dorsal and ventral. In human vision, information passes from the low-level perceptual areas at the back of the brain via the ventral stream to posterior temporal areas, to determine *what* an object is, and via the dorsal stream to parietal cortex and motor areas, to determine *where* the object is and *how* it can be manipulated. Applying this same dual-stream hypothesis to language processing, the ventral stream through the temporal lobe is responsible for word semantics (i.e., what), whereas the dorsal stream links motor areas of the brain (including the articulatory network in the posterior inferior frontal gyrus) with auditory and sensorimotor areas of the brain.

Hagoort (2005) proposes a related model that focuses on the neural mechanisms for the unification of language (which covers both semantic composition and syntactic operations), which enables the generation of composed meaning, considered by many theories to be the central defining feature of human language competence. The MUC model consists of three functions:

Memory: Recalling the meaning of a word, lexical access. The temporal cortex and the inferior parietal cortex are involved in the memory process of Hagoort's model.

Unification: Integrating the retrieved meaning of a word with the meaning representation calculated with the context leading up to that word. This

includes extralinguistic sources of meaning like gesture and gender of speaker. This processing resides in the left inferior frontal cortex, with Brodmann areas BA 47 and BA 45 specialized for semantic unification, and neighboring areas BA 45 and BA 44 (Broca's area) specialized for syntax (see Figure 3.1) (Hagoort, 2014).

Control: Governing the actions required for language, like taking turns during a conversation. Control requires the dorsolateral prefrontal cortex, anterior cingulate cortex (ACC) and the parts of the parietal cortex that govern attention.

The model of Kuperberg (2007) is primarily based on EEG evidence and does not make strong claims on localization. Despite this, it does posit two parallel processes in language (semantic memory and semantic combination), which have similarities to the dual-stream model and mirror two of the three components of the MUC model. Under this model, the P600 is due to the continued analysis required if the output of this combinatorial stream is incongruent with the output of the predictions of the semantic stream. The combinatorial stream processes two types of constraints: morphosyntactic information, which is used to resolve syntax errors, and semantic-thematic information, which can influence the N400 because it operates in parallel with the semantic memory stream. Processing of this constraint may continue after the N400 window if more combinatory analysis is needed.

As we have seen in this section, neuroimaging studies have concentrated on seeking answers about the neural organization of the language faculty, and the brain's processing of linguistic input. See Figure 3.1 for a summary. Mostly absent is the consideration of more specific details of languages, such as the representation of words, phrases, sentences, and the structures that underlie them. In the following sections we describe recent studies that combine computational modeling of language with recordings of brain activities to examine the finer grain of languages.

3.3 Decoding Words in the Brain

As mentioned in the previous section, architectural theories of language processing see the temporal lobe as central to the generic retrieval and processing of words. Less work has been devoted to characterizing the representation and processing of particular words or word classes.

There is a great deal literature on the representation of classes of *objects*, represented using images. For instance there is well-documented specialization for semantic class in the inferior temporal cortex, with areas of the fusiform gyrus in particular being differentially sensitive to living and nonliving things

(Martin et al., 1996). In the vision science community this is considered “high-level vision,” where the low-level visual input has been abstracted away to the extent that these parts of the brain encode something about the meaning and content of an image (Connolly et al., 2012; Carlson et al., 2013). An obvious question then is whether the representations in these brain areas are indeed visual, or are rather amodal. One study demonstrated similar patterns of activity, specific to semantic categories in these “visual” areas, in congenitally blind participants (Mahon et al., 2009), suggesting that this area (conveniently located next to language semantic areas) may be the locus of amodal (and perhaps symbolic) meaning. A series of subsequent studies found commonalities in brain activity patterns evoked by words, and by their corresponding pictures, with left posterior temporal areas emerging as key to amodal decoding (Shinkareva et al., 2011; Devereux et al., 2013; Fairhall and Caramazza, 2013; Simanova et al., 2014).²

In recent years, embodied theories of meaning (Barsalou and Wiemer-Hastings, 2005) have challenged this classical position. A particularly influential study looked for evidence of the sensory/motor coding of common concepts. Pulvermüller (2005) demonstrated that passive reading of physical action verbs caused increases in brain areas responsible for controlling the corresponding body parts. For instance, reading the verb *lick* caused activation in the vicinity of the face sensory/motor area, *pick* in the hand area, and *kick* in the leg area. In addition to the strong claim that meaning was inherently modal, this added to the evidence that conceptual information, whatever its content, was coded in a distributed fashion across the brain (Haxby et al., 2001; Marslen-Wilson et al., 2001; Martin, 2007), and is also consistent with the position that the brain uses both amodal and modality-specific representations when processing semantics (Fernandino et al., 2016; Handjaras et al., 2016).

Most studies of this type have an inherent practical limitation. Conventional experiment designs, and the cost involved in collecting brain data, both conspire to restrict us to small numbers of stereotyped concept types (such as the overstudied animals and tools [e.g., Murphy et al., 2011]). Because an adult vocabulary consists of tens of thousands of entries, it is hard to see how such an approach could attempt to provide a comprehensive account of our mental lexicon.

A pioneering study by Mitchell et al. (2008) provides an alternative paradigm for studying language and other types of higher cognition. Rather than characterizing a cognitive faculty (in this case, language) in terms of a small number of constructed conditions, it attempts to describe the fine grain and the breadth

² Simanova et al. (2014) stands out by using spoken words and associated sounds (e.g., a dog’s bark) in addition to the pictorial and written word stimuli used in most other studies.

of language. It takes advantage of models of language from computational linguistics and uses these as an intermediate feature description to establish the mapping between brain recordings and the stimuli and tasks that evoke them.

Mitchell set out to computationally describe the brain activity associated with single words and concepts. This posited, possibly distributed, pattern of brain activity is termed a *neural signature*. Recordings of brain activity are noisy, and practical limitations prevent us from collecting the large amounts of data that would be needed to directly discover the neural signature for a single word. Instead, an intermediate model describes the similarities and dissimilarities among the concepts of interest in terms of semantic dimensions. Whereas linguistics and lexicography have historically relied upon binary and categorial features (Katz and Fodor, 1963; Fillmore, 1982; Miller et al., 1990), more recently word spaces and word vectors (embeddings) have become broadly used as an empirically grounded description of word meaning (Lund et al., 1995; Landauer and Dumais, 1997; Collobert and Weston, 2008; Baroni and Lenci, 2010). Mitchell used a simplified vector space model (VSM), based on

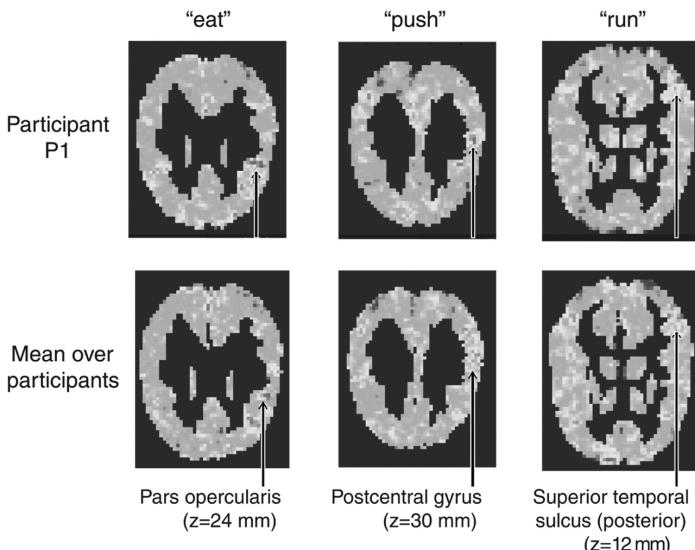


Figure 3.4 Several slices from fMRI images showing the *learned* proportion of brain activation that can be associated with a particular verb from the set of twenty-five verbs used in Mitchell et al. (2008). Note that verbs map onto areas of the brain previously shown to be associated with related semantic information. Figure courtesy of Mitchell et al. (2008).

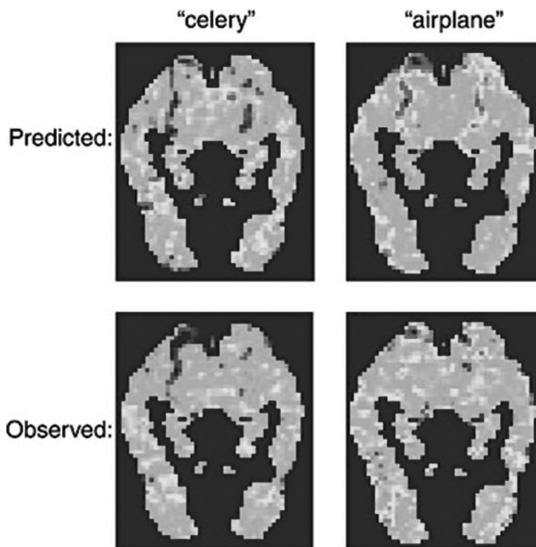


Figure 3.5 The predicted (*top row*) and observed (*bottom row*) brain activation for a particular person while reading the word “celery” or “airplane” (*left and right columns, respectively*). Although the brain images are not perfect matches, it is clear that the observed activity for celery is closer to the predicted activation for celery than airplane, and similarly the observed airplane activation is more similar to the predicted airplane activation. Figure adapted from Mitchell et al. (2008).

co-occurrence statistics with verbs encoding sensory-motor interaction, according to a large set of web *n*-gram statistics.

On the basis of this a regularized regression model was trained to find mappings between individual model features and the recordings of fMRI activity, in the process identifying components of brain activity with a semantic interpretation (Just et al., 2010). A side effect of discovering neural signatures for these semantic components was that such components can be recombined in novel ways to estimate the neural signature for unseen concepts (that is, words or images that were not part of the computational training set). Examples of semantic components discovered in brain activity are shown in Figure 3.4, whereas Figure 3.5 compares several observed neural signatures to the estimated reconstruction for those concepts (when they were unseen by the algorithm).

This made it possible to guess, based on a person’s brain activity, which concept was being processed. Of course it is conceivable that such brain decoding performance could be driven by low-level confounds with semantics:

for instance artifacts tend to be more angular and less visually textured than natural kinds. But the use of a particular set of intermediate features, encoding semantics rather than perceptual characteristics, suggests that semantics is the driving factor. A follow-on study also demonstrated that decoding works across languages, e.g., training on fMRI data from English speakers, and decoding what concept a Portuguese speaker is thinking about (Buchweitz et al., 2012).

Follow-on studies established that this analytical paradigm worked with other semantic feature spaces, including crowdsourced semantic judgments (Palatucci, 2011), psychological feature norms (Chang et al., 2011), association norms (Akama et al., 2015), structured taxonomies (Jelodar et al., 2010), Wikipedia definitions (Pereira et al., 2010), and a broader range of word space embeddings (Devereux and Kelly, 2010; Murphy et al., 2012b; Bullinaria and Levy, 2013; Fyshe et al., 2014). Similar analyses have been performed with EEG (Murphy et al., 2009; Simanova et al., 2010) and MEG (Sudre et al., 2012); using videos as stimuli in place of words or images (Huth et al., 2012); for other classes of words beyond the small number of concrete categories used in the original experiments (Anderson et al., 2014); and reversing the direction of inference, so that semantic features can be “read out” of brain images in an approximate fashion (Pereira et al., 2011). Across all of these studies of words in the brain, it is noted that semantic representations are distributed throughout the cortex, and there is no singular locus of lexical meaning.

3.4 Phrases in the Brain

Once we can detect the neural signatures of individual words, we can study how they combine in the brain to form composed phrases. There have been several studies on adjective phrase processing by normal adults, including MEG studies that have implicated right and left anterior temporal lobes (RATL and LATL) as well as ventromedial prefrontal cortex (vmPFC). Adjective-noun pairs elicit increased neural activity when compared with word lists or nonwords paired with nouns, with activity significantly higher in LATL (~ 200 ms), vmPFC ($\sim 300\text{--}500$ ms), and RATL (~ 200 and $300\text{--}400$ ms) (Bemis and Pylkkänen, 2011). When comparing a compositional picture-naming task with a noncompositional picture-naming task, Bemis and Pylkkänen (2013) found differences in the magnitude of activation in LATL. Bemis and Pylkkänen hypothesize that, due to the timing of these effects, the activity in vmPFC is related to semantic processes, and that LATL activity could be due to the either the syntactic or semantic demands of composition.

Semantic composition has also been studied using word vector space models. The computational linguistics community has proposed several methods of

combining word vectors to produce an estimate of the meaning of a phrase, including adding or multiplying the vectors together (Mitchell and Lapata, 2010; Erk, 2012). These studies have spurred several brain-imaging experiments that look for the composed semantic representation in the brain.

Adjective-noun composition in fMRI was explored by Chang et al. (2009) with twelve adjective-noun pairs and corpus-derived vectors composed of verb co-occurrence statistics, inspired by Mitchell et al. (2008). They showed that, in terms of R^2 (regression coefficient of determination, or variation explained), a multiplicative model of composition outperformed an additive composition model, and also the adjective or the noun's semantic vector. However, in terms of ranking the predicted brain activation under the learned model by distance to the true brain activation (as in Mitchell et al.), the additive, multiplicative, and noun-only model performed similarly. Fyshe et al. (2013) explored using these composed vector representations to decode the phrase from MEG data. They showed that a more complex function, "dilation," gave better decoding of phrases from MEG data. Dilation emphasizes the dimensions of the noun that are shared with the adjective, which is a plausible metaphor for the composition of some adjective-noun phrases.

Baron and Osherson (2011) studied the semantic composition of adjective-noun phrases using fMRI. Here, the stimuli was not linguistic, but rather was the faces of young or old males (boys and men) and young or old females (girls and women). In the fMRI scanner, the faces were presented in the same order for several minutes (time block). For each time block within the experiment, subjects were given a category (e.g., girl) and were asked to determine whether each of the stimuli faces was a member of that category. Thus, for each block, the face stimuli was constant, and only the concepts being matched (e.g., young and female) differed. Thus, any differences in activation can be attributed only to the matching task, and not to the stimuli. Baron and Osherson then created conceptual maps by learning models to predict brain activity based on the age (young or old) or gender of the matching task. They found that the brain activation of a composed concept (e.g., young male) could be estimated by the multiplication or addition of adjective (e.g., young) and noun (e.g., boy) brain activation maps. Areas of the brain that could be approximated well with an additive function were widespread and covered frontal, parietal, occipital, and temporal lobes, whereas the multiplicative function was useful for predicting just to the LATL.

How do the experimental results for semantic composition relate to the models of semantic unification previously discussed? If the syntactic form is held constant (as in adjective-noun phrases), the brain processes for syntactic combination are identical. However, when the semantics of the phrase changes, the semantic retrieval/memory and unification processes will also change, resulting in differential brain activity.

In Bemis and Pylkkänen's work, the semantic content of the words is constant, but the task differs (combining words into phrases or not). Their findings show increased activation in LATL, RATL, and vmPFC, which implies that the combinatorial processes of adjective-noun composition are at least partially handled in these areas. This is consistent with the local structure-building process in Figure 3.1, involving ventral pathway II. Hickok and Poeppel (2007) also hypothesize that the anterior temporal lobe is involved in composition, although they localize it to a slightly more medial temporal location.

Recall from our discussion of single words in the brain that semantics is distributed throughout the cortex. This finding implies that the semantic portion of semantic composition will likely occur in many places in the brain, even if composition is mediated by areas of the temporal lobe. This is congruent with the additive model of Baron and Osherson (2011). Perhaps the temporal lobe (indicated by multiple studies for the syntactic processes of composition) acts like the conductor of an orchestra, and each of the distributed semantic areas is an instrument. Signals are sent by the conductor to raise or lower particular elements of the orchestra, or to cause specific areas to play in synchrony. This is a metaphor for the way the brain could encode changes in semantics due to composition, bringing the activation of brain areas up or down, or causing areas to work in synchrony to encode meaning altered by context.

3.5 Stories in the Brain

Story processing is emerging as a new, more ecologically valid way to study the finer grain of language in the brain. Using authentic materials allows us to simultaneously examine the multiple levels of processing engaged in language comprehension. By moving to the narrative we can also investigate discourse factors (e.g., keeping track of events, characters' perspectives, reader response) beyond the levels of syntax and semantics seen in sentence studies.

Mason and Just (2006) review many studies involved in processing narrative, and identify networks involved in story understanding: a coarse semantic processing network (in the right temporal lobe), a topical coherence monitoring network in the bilateral frontal lobes, a text integration network in the anterior temporal and inferior frontal cortices. They also identify networks that process narrative information: a "protagonist's perspective interpreter network" in the right superior temporal cortex and the medial frontal lobes, and an imagery network in the bilateral intraparietal sulcus. In Speer et al. (2009), changes along different narrative dimensions were manually annotated (e.g., changes in goal, time, character identity, or location). Different regions in the brain had their activity correlated with these dimensions. Notably, the change in character identity also correlated with posterior lateral and medial frontal activity along with other temporal regions. This experiment and others mentioned in

Mason and Just (2006) mostly used hand-labeled features instead of computational language models used for text annotation. Story comprehension tasks have also been used to examine how the implied affect of individual words relates to that of phrases and larger passages (Hsu et al., 2015).

Other studies of story in fMRI focus on language processing instead of narrative structure, and many of them focus specifically on syntactic processing, and the cognitive load imposed by structures of varying complexity. In Bachrach (2008), tailored stories are used to auditorily present complex syntactic structures with higher frequency than average. Multiple measures of syntactic structure are computed and found to correlate with the activity of multiple brain regions in the temporal and inferior frontal cortices, including the left anterior temporal lobes. Theory-of-mind features were also used, and, once again, found to be correlated with the activity in the posterior temporal cortex. In Brennan et al. (2010), syntactic load is measured by building a parse tree of every sentence and computing the tree depth of each word. This syntactic feature also predicts the activity in the anterior temporal lobe, suggesting that the anterior temporal lobe is involved in the structural aspect of sentence composition as well as the semantic aspect, or that the two might be hard to disentangle. Mechanisms of memory, anticipation, and information structure are also being studied intensively (Hale et al., 2015; van Schijndel et al., 2015; Frank et al., 2015).

In Wehbe et al. (2014b), subjects read a chapter of J. K. Rowling's fantasy novel *Harry Potter and the Sorcerer's Stone* in the fMRI scanner in rapid serial visual presentation mode. Computational language models were used to label the words of the chapter. A semantic feature space of the words was constructed using non-negative sparse embedding (Murphy et al., 2012a), and syntactic features were obtained by identifying the part of speech and grammatical role of the words using the MaltParser (Nivre et al., 2007). Multiple regions spanning the bilateral temporal cortices were found to represent syntax or semantics, and sometimes both, hinting to the possibility that syntax and semantics might be nondissociated concepts. Other hand-labeled features were identified that characterized narrative components such as the presence of different story characters (which corresponded to activity in classical theory-of-mind areas) and their physical motions (which corresponded to activity in regions also activated during the perception of biological motion). The results are shown in Figure 3.6.

Huth et al. (2016) had subjects listen to hours of narrated stories (from *The Moth Radio Hour* series of podcasts). The authors built a VSM based on word co-occurrences with a set of nine hundred and eighty-five frequently used English words. A generative model was estimated that predicts brain activity as a function of this semantic VSM, and it was able to predict a considerable portion of the variance of activity across all regions typically referred to as the semantic system, spanning most of the temporal cortices, parts of the parietal

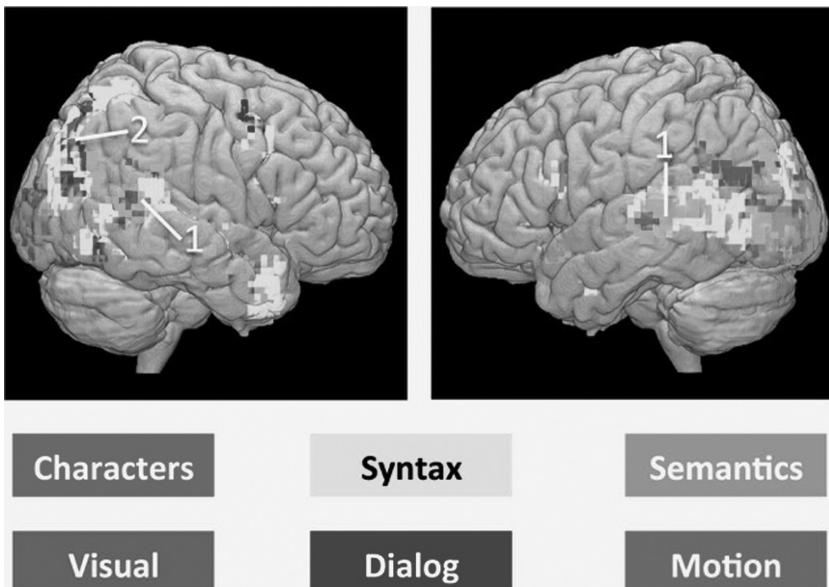


Figure 3.6 Story reading brain map, adapted from Wehbe et al. (2014b). “Results obtained by [...] a generative model, showing where semantic, discourse, and syntax information is encoded by neural activity. Note this model identifies not just where language processing generates neural activity, but also what types of information are encoded by that activity. Each voxel location represents the classification done using a cube of $5 \times 5 \times 5$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that cube are used. Voxel locations are colored according to the feature set that can be used to yield significantly higher than chance accuracy. Light green regions, marked with (1), are regions in which using either semantic or syntactic features leads to high accuracy. Dark gray regions, marked with (2), are regions in which using either dialog or syntactic features leads to high accuracy.”

cortices, and the frontal cortices. Importantly, and in conjunction with the previously mentioned studies, the uncovered semantic representation was highly bilateral. Furthermore, the authors constructed an atlas of semantic representations by combining data across all their subjects using a model that accounts for individual variations. They were able to identify a large set of semantically selective areas that each encode a constrained set of concepts. One of their findings is that the bilateral posterior temporal cortex is highly responsive to words related to social interaction. Previously we saw this region activated by tasks related to keeping track of the protagonist perspective.

Another method for studying language in a naturalistic scenario is to find voxels that are highly correlated across subjects in various conditions. Namely,

Lerner et al. (2011) contrasted intersubject voxel correlations when listening to paragraphs, scrambled sentences, scrambled words, and backward speech. The authors found the activity was consistent among subjects in each condition in a set of increasingly larger brain regions that are hierarchically organized: in the simplest condition (backward speech), the activity was consistent mostly in the primary auditory cortex. As the temporal integration window became longer when moving to words, sentences, and paragraphs, the consistent region became larger and encompassed a larger and larger part of the middle temporal cortex, eventually spreading to the posterior temporal cortex, the temporoparietal junction, and the inferior frontal gyri. This study highlights the importance of studying the difference between the representation of the combined meaning of words and the representation of these words in isolation.

To investigate this question, the same *Harry Potter* experiment was performed in MEG by Wehbe et al. (2014a). This experiment studied the representations of the properties of a word versus the properties of the context that preceded it, and tried to identify the different stages of continuous meaning construction when subjects read a text. The researchers used a recurrent neural network language model (Mikolov, 2012) to obtain feature representations for the context of a word (computed before the word appeared) and the features of

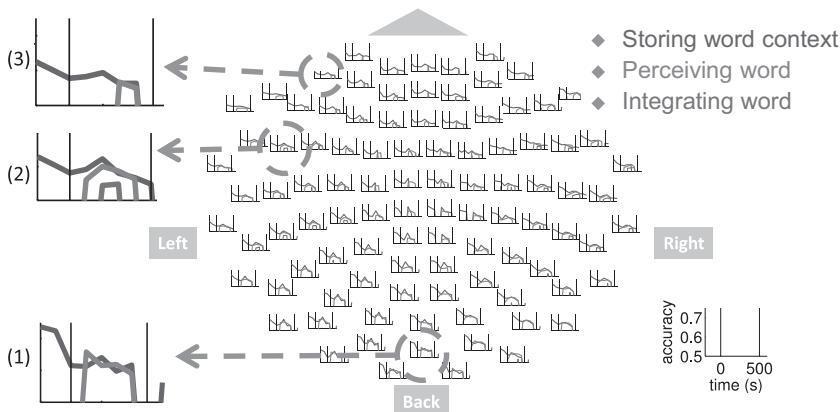


Figure 3.7 Adapted from Wehbe et al. (2014a). Time-line of word integration across the MEG helmet. For each of the one hundred and two locations the average accuracy at that location is plotted versus time. The axes are defined in the rightmost, empty plot. Accuracy should be seen as an indicator for the underlying process. Three plots have been magnified to show the increasing delay in processing the incoming word as information travels through the brain from the visual cortex (1) to the temporal (2) and frontal (3) cortices. Furthermore, the integration process is shown to occur in the left and right temporal cortices, after the word perception begins. The word context is maintained before the word is shown, and is gradually updated as the word is perceived.

that word. The recurrent neural network language model iteratively combines consecutive words and predicts the incoming word while maintaining a hidden vector representation of context. The model was run on the words of the chapter and the context vector, as well as the vector specifying the properties of the current word, and a measure of how surprising that word was given its context was extracted at each step. These vectors were then used to predict brain activity. The results, shown in Figure 3.7, reveal that context is more predictive of brain activity than the properties of the incoming word, hinting that more brain activity might be involved in representing context. Furthermore, the results include a suggested time line of how the brain updates its context representation. They also demonstrate the incremental perception of every new word starting early in the visual cortex, moving next to the temporal lobes, and finally to the frontal regions. Lastly, the results suggest that the integration process occurs in the temporal lobes after the new word has been perceived.

3.6 Summary

Brain activity data is the most direct record we have of the psychological states and processes that underlie language function. In this chapter, we reviewed earlier work that studied neuroimaging data using the tools of experimental psychology. This approach has provided new insights into the macrofunction and architecture of the language faculty, but can be limited in the generality and detail it can describe. Because brain data is difficult to obtain, statistical and machine learning methods that rely on large amounts of data have limited traction. This means that computational linguistics is an ideal and complementary tool, leveraging the huge quantities of naturalistic text available on the web to build detailed models and instantiate fine-grained theories. The use of computational features of text as an intermediate description enables models of greater generality that may draw conclusions beyond the limited set of stimuli that can be presented in a single experiment.

Naturalistic experimental tasks can also play a role, again letting computational models resolve the multifactorial complexity of human language comprehension. Authentic texts as stimuli do present considerable challenges to analysis (e.g., the systematic confounding among syntax and semantics, and Zipfian distribution of words and phrase categories [see, e.g., Hasson and Egidi, 2013]), for which methodological solutions continue to emerge. At the same time they are excellent from the point of view of engaging participants in authentic language processing, and are more representative of real-world language experience than many hand-tailored materials. This more holistic approach to understanding language processing is also consistent with an increasing tendency in neuroscience to understand brain activity in terms of networks and interactions among functional units.

Looking to the future, we expect to see continued progress, driven by larger and more varied datasets, and more powerful learning algorithms (as in computational linguistics, deep learning methods are beginning to impact neuroscience, e.g., Koyamada et al., 2014; Zheng et al., 2014). As methods and data improve, we hope to gain greater insight into the fundamental questions of linguistics, such as the universality of representations (Zinszer et al., 2015), the (in)dependence of syntax and semantics, and question of how language knowledge is encoded (Handjaras et al., 2016) and interacts with real-world and procedural knowledge.

Data sharing in neuroscience is increasing, driven both from grassroots (Yarkoni et al., 2010) and by the policies of governments and other funders. Another area of rapid change is analyses that make use of multiple modalities of data, combining recordings of brain activity with textual data (Fyshe et al., 2014), eye-gaze tracking (Desai et al., 2016; Henderson et al., 2016), and collections of natural images (Khaligh-Razavi and Kriegeskorte, 2014; Clarke et al., 2015; Anderson et al., 2015). And finally, advances in analysis may help our understanding of individual variation in language processing (see, e.g., Charest et al., 2014), outside the lab as well (Kidmose et al., 2013).

References

- Akama, Hiroyuki, Miyake, Maki, Jung, Jaeyoung, and Murphy, Brian. 2015. Using graph components derived from an associative concept dictionary to predict fMRI neural activation patterns that represent the meaning of nouns. *PLoS one*, **10**(4), e0125725.
- Anderson, Andrew J., Murphy, Brian, and Poesio, Massimo. 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, **26**(3), 658–81.
- Anderson, Andrew James, Bruni, Elia, Lopopolo, Alessandro, Poesio, Massimo, and Baroni, Marco. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, **120**, 309–322.
- Bachrach, Asaf. 2008. “Imaging Neural Correlates of Syntactic Complexity in a Naturalistic Context.” PhD thesis, Massachusetts Institute of Technology.
- Baron, Sean G., and Osherson, Daniel. 2011. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, **55**(4), 1847–52.
- Baroni, Marco, and Lenci, Alessandro. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.
- Barsalou, Lawrence W., and Wiemer-Hastings, Katja. 2005. Situating abstract concepts. Chap. 7, pages 129–163, in Pecher, D., and Zwaan, R. (eds), *Grounding Cognition: The Role of Perception and Action in Memory Language and Thinking*. Cambridge, UK: Cambridge University Press.
- Bear, Mark F., Connors, Barry W., and Paradiso, Michael A. 2007. *Neuroscience: Exploring the Brain*. 3rd ed. Baltimore: Lippincott Williams & Wilkins.

- Bemis, D. K., and Pylkkänen, L. 2013. Basic linguistic composition recruits the left anterior temporal lobe and left angular gyrus during both listening and reading. *Cerebral Cortex*, **23**(8), 1859–73.
- Bemis, Douglas K., and Pylkkänen, Liina. 2011. Simple composition: A magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of Neuroscience*, **31**(8), 2801–14.
- Brennan, Jonathan, and Pylkkänen, Liina. 2012. The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, **60**(2), 1–10.
- Brennan, Jonathan, Nir, Yuval, Hasson, Uri, Malach, Rafael, Heeger, David J, and Pylkkänen, Liina. 2010. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, **120**(2), 163–73.
- Buchweitz, Augusto, Shinkareva, Svetlana V, Mason, Robert A, Mitchell, Tom M, and Just, Marcel Adam. 2012. Identifying bilingual semantic neural representations across languages. *Brain and Language*, **120**(3), 282–9.
- Bullinaria, John A., and Levy, Joseph P. 2013. Limiting factors for mapping corpus-based semantic representations to brain activity. *PloS one*, **8**(3), e57191.
- Carlson, Thomas, Tovar, Da, Alink, Arjen, and Kriegeskorte, Nikolaus. 2013. Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, **13**(10), 1–19.
- Chang, Kai-min, Cherkassky, Vladimir L., Mitchell, Tom M., and Just, Marcel Adam. 2009. Quantitative modeling of the neural representation of adjective-noun phrases to account for fMRI activation. Pages 638–646 of: *Proceedings of the Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*.
- Chang, Kai-min Kevin, Mitchell, Tom, and Just, Marcel Adam. 2011. Quantitative modeling of the neural representation of objects: how semantic feature norms can account for fMRI activation. *NeuroImage*, **56**(2), 716–27.
- Charest, Ian, Kievit, Rogier A., Schmitz, Taylor W., Deca, Diana, and Kriegeskorte, Nikolaus. 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(40), 14565–70.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- Clarke, Alex, Devereux, Barry J., Randall, Billi, and Tyler, Lorraine K. 2015. Predicting the time course of individual objects with MEG. *Cerebral Cortex*, **25**(10), 3602–3612.
- Collobert, Ronan, and Weston, Jason. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th International Conference on Machine Learning*, 160–167.
- Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., Abdi, H., and Haxby, J. V. 2012. The Representation of Biological Classes in the Human Brain. *Journal of Neuroscience*, **32**(8), 2608–2618.
- DeLong, Katherine A., Urbach, Thomas P., and Kutas, Marta. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, **8**(8), 1117–1121.
- Desai, Rutvik H., Choi, Wonil, Lai, Vicky T., and Henderson, John M. 2016. Toward semantics in the wild: Activation to manipulable nouns in naturalistic reading. *Journal of Neuroscience*, **36**(14), 4050–4055.

- Devereux, Barry, and Kelly, Colin. 2010. Using fMRI activation to conceptual stimuli to evaluate methods for extracting conceptual representations from corpora. In: Murphy, Brian, Korhonen, Anna, and Chang, Kevin Kai-Min (eds), *1st Workshop on Computational Neurolinguistics*.
- Devereux, Barry J., Clarke, Alex, Marouchos, Andreas, and Tyler, Lorraine K. 2013. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of Neuroscience*, **33**(48), 18906–16.
- Dronkers, N. F., Plaisant, O., Iba-Zizen, M. T., and Cabanis, E. A. 2007. Paul Broca's historic cases: High resolution MR imaging of the brains of Leborgne and Lelong. *Brain*, **130**(5), 1432–1441.
- Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, **6**(10), 635–653.
- Fairhall, Scott L., and Caramazza, Alfonso. 2013. Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, **33**(25), 10552–10558.
- Fedorenko, Evelina, Nieto-Castañon, Alfonso, and Kanwisher, Nancy. 2012. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, **50**(4), 499–513.
- Fernandino, Leonardo, Binder, Jeffrey R., Desai, Rutvik H., Pendl, Suzanne L., Humphries, Colin J., Gross, William L., Conant, Lisa L., and Seidenberg, Mark S. 2016. Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex*, **26**(5), 2018–2034.
- Fillmore, Charles J. 1982. Frame semantics. Pages 111–138 of: Korea, Linguistic Society (ed), *Linguistics in the Morning Calm*. Seoul: Hanshin.
- Fodor, Jerry A. 1983. *Modularity of Mind*. Cambridge, MA: MIT Press.
- Frank, Stefan L., Otten, Leun J., Galli, Giulia, and Vigliocco, Gabriella. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, **140**, 1–11.
- Friederici, Angela D. 2002. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, **6**(2), 78–84.
- Friederici, Angela D. 2011. The brain basis of language processing: From structure to function. *Physiological Reviews*, **91**(4), 1357–92.
- Fyshe, Alona. 2015. “Corpora and Cognition: The Semantic Composition of Adjectives and Nouns in the Human Brain,” PhD thesis, Carnegie Mellon University.
- Fyshe, Alona, Talukdar, Partha, Murphy, Brian, and Mitchell, Tom. 2013. Documents and Dependencies: An Exploration of Vector Space Models for Semantic Composition. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 84–93.
- Fyshe, Alona, Talukdar, Partha P. P., Murphy, Brian, and Mitchell, Tom M. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. Pages 489–499 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 1.
- Gibson, Edward, and Fedorenko, Evelina. 2010. The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, **28**(1), 1–37.
- Hagoort, Peter. 2005. On Broca, brain, and binding: A new framework. *Trends in Cognitive Sciences*, **9**(9), 416–23.

- Hagoort, Peter. 2008. The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **363**(1493), 1055–69.
- Hagoort, Peter. 2014. Nodes and networks in the neural architecture for language: Broca's region and beyond. *Current Opinion in Neurobiology*, **28C**(Jul), 136–141.
- Hagoort, Peter, Hald, Lea, Bastiaansen, Marcel, and Petersson, Karl Magnus. 2004. Integration of word meaning and world knowledge in language comprehension. *Science*, **304**(5669), 438–41.
- Hale, John T., Lutz, David E., Luh, Wen-ming, Brennan, Jonathan R., and Arbor, Ann. 2015. Modeling fMRI time courses with linguistic structure at various grain sizes. Pages 89–97 of: *Proceedings of the Sixth Workshop on Cognitive Modeling and Computational Linguistics*.
- Hamalainen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. 1993. Magnetoencephalography theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, **65**(2).
- Handjara, Giacomo, Ricciardi, Emiliano, Leo, Andrea, Lenci, Alessandro, Cecchetti, Luca, Cosottini, Mirco, Marotta, Giovanna, and Pietrini, Pietro. 2016. How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, **135**, 232–242.
- Hansen, Peter, Krriegelbach, Morten, and Salmelin, Riitta. 2010. *MEG: An Introduction to Methods*. Boston: Oxford University Press.
- Hasson, Uri, and Egidi, Giovanna. 2013. What are naturalistic comprehension paradigms teaching us about language? Chap. 11, pages 228–255 of: Willems, Roel M (ed), *Cognitive Neuroscience of Natural Language Use*. Cambridge University Press.
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, **293**(5539), 2425–2430.
- Henderson, John M., Choi, Wonil, Lowder, Matthew W., and Ferreira, Fernanda. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *NeuroImage*, **132**, 293–300.
- Hickok, Gregory, and Poeppel, David. 2004. Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, **92**(1–2), 67–99.
- Hickok, Gregory, and Poeppel, David. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**(May), 393–402.
- Hoeks, John C. J., Stowe, Laurie A., and Doedens, Gina. 2004. Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research*, **19**(1), 59–73.
- Hsu, Chun-Ting, Jacobs, Arthur M., Citron, Francesca M. M., and Conrad, Markus. 2015. The emotion potential of words and passages in reading Harry Potter - An fMRI study. *Brain and Language*, **142**, 96–114.
- Huth, Alexander G., Nishimoto, Shinji, Vu, An T., and Gallant, Jack L. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, **76**(6), 1210–24.

- Huth, Alexander G., de Heer, Wendy A., Griffiths, Thomas L., Theunissen, Frédéric E., and Gallant, Jack L. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, **532**(7600), 453–458.
- Im, Chang Hwan, Gururajan, Arvind, Zhang, Nanyin, Chen, Wei, and He, Bin. 2007. Spatial resolution of EEG cortical source imaging revealed by localization of retinotopic organization in human primary visual cortex. *Journal of Neuroscience Methods*, **161**, 142–154.
- Jelodar, Ahmad Babaeian, Alizadeh, Mehrdad, and Khadivi, Shahram. 2010. WordNet Based Features for Predicting Brain Activity associated with meanings of nouns. Pages 18–26 of: Murphy, Brian, Korhonen, Anna, and Chang, Kevin Kai-Min (eds), *1st Workshop on Computational Neurolinguistics*.
- Just, Marcel Adam, Cherkassky, Vladimir L., Aryal, Sandesh, and Mitchell, Tom M. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one*, **5**(1), e8622.
- Katz, Jerrold J., and Fodor, Jerry A. 1963. The structure of a semantic theory. *Language*, **2**(39), 170–210.
- Khaligh-Razavi, Seyed-Mahdi, and Kriegeskorte, Nikolaus. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, **10**(11), e1003915.
- Kidmose, Preben, Looney, David, Ungstrup, Michael, Rank, Mike Lind, and Mandic, Danilo P. 2013. A study of evoked potentials from ear-EEG. *IEEE Transactions on Biomedical Engineering*, **60**(10), 2824–2830.
- Koyamada, Sotetsu, Shikauchi, Yumi, Nakae, Ken, and Ishii, Shin. 2014. Construction of Subject-independent Brain Decoders for Human fMRI with Deep Learning. Pages 60–68 of: *The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)*.
- Kuperberg, Gina R. 2007. Neural mechanisms of language comprehension: challenges to syntax. *Brain Research*, **1146**(May), 23–49.
- Kutas, M., and Hillyard, S. SA. 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, **207**(4427), 203–5.
- Kutas, Marta, and Federmeier, Kara D. 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, **62**(jan), 621–47.
- Landauer, T., and Dumais, S. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**(2), 211–240.
- Lerner, Yulia, Honey, Christopher J., Silbert, Lauren J., and Hasson, Uri. 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of Neuroscience*, **31**(8), 2906–2915.
- Lund, K., Burgess, C., and Atchley, R. 1995. Semantic and associative priming in high dimensional semantic space. Pages 660–665 of: *Proceedings of the 17th Cognitive Science Society Meeting*.
- Mahon, Bradford Z., Anzellotti, Stefano, Schwarzbach, Jens, Zampini, Massimiliano, and Caramazza, Alfonso. 2009. Category-specific organization in the human brain does not require visual experience. *Neuron*, **63**(3), 397–405.
- Marslen-Wilson, William, and Tyler, Lorraine Komisarjevsky. 1980. The temporal structure of spoken language understanding. *Cognition*, **8**(1), 1–71.

- Marslen-Wilson, William, Tyler, Lorraine Komarjevsky, and Moss, Helen E. 2001. Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, **5**(6), 244–252.
- Martin, Alex. 2007. The representation of object concepts in the brain. *Annual Review of Psychology*, **58**(1), 25–45.
- Martin, Alex, Wiggs, Cheri L, Ungerleider, Leslie G, and Haxby, James V. 1996. Neural correlates of category-specific knowledge. *Nature*, **379**(Feb), 649–652.
- Mason, R. A., and Just, M. A. 2006. Neuroimaging contributions to the understanding of discourse processes. Pages 765–799 of: Traxler, M., and Gernsbacher, M. A. (eds), *Handbook of Neuropsychology*. Amsterdam: Elsevier.
- McCandliss, Bruce D., Cohen, Laurent, and Dehaene, Stanislas. 2003. The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, **7**(7), 293–299.
- Mikolov, Tomas. 2012. “Statistical Language Models Based on Neural Networks. PhD thesis, Czech Republic: Brno University of Technology”.
- Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, and Miller, Katherine. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, **3**(4), 235–244.
- Mitchell, Jeff, and Lapata, Mirella. 2010. Composition in distributional models of semantics. *Cognitive Science*, **34**(8), 1388–429.
- Mitchell, Tom M., Shinkareva, Svetlana V., Carlson, Andrew, Chang, Kai-Min, Malave, Vicente L., Mason, Robert A., and Just, Marcel Adam. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, **320**(5880), 1191–5.
- Murphy, Brian, Baroni, Marco, and Poesio, Massimo. 2009. EEG responds to conceptual stimuli and corpus semantics. Pages 619–627 of: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.
- Murphy, Brian, Poesio, Massimo, Bovolo, Francesca, Bruzzone, Lorenzo, Dalponte, Michele, and Lakany, Heba. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, **117**(1), 12–22.
- Murphy, B., Talukdar, P., and Mitchell, T. 2012a. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In: *International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India.
- Murphy, Brian, Talukdar, Partha, and Mitchell, Tom. 2012b. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. Pages 114–123 of: *First Joint Conference on Lexical and Computational Semantics (*SEM)*.
- Nivre, J., Hall, J., Nilsson, J., Chaney, A., Eryigit, G., Kubler, S., Marinov, S., and Marsi, E. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95.
- Özyürek, Asli, Willem, Roel M., Kita, Sotaro, and Hagoort, Peter. 2007. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, **19**(4), 605–616.
- Palatucci, Mark M. 2011. “Thought Recognition: Predicting and Decoding Brain Activity Using the Zero-Shot Learning Model.” PhD thesis, Carnegie Mellon University.
- Pereira, Francisco, Botvinick, Matthew, and Detre, Greg. 2010. Learning semantic features for fMRI data from definitional text. In: Murphy, Brian, Korhonen, Anna, and Chang, Kevin Kai-Min (eds), *1st Workshop on Computational Neurolinguistics*.

- Pereira, Francisco, Detre, Greg, and Botvinick, Matthew. 2011. Generating text from functional brain images. *Frontiers in Human Neuroscience*, **5**(August), 1–11.
- Pulvermüller, Friedemann. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, **6**, 576–582.
- Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**(3), 372–422.
- Salmelin, Riitta. 2007. Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology*, **118**(2), 237–54.
- Shinkareva, Svetlana V., Malave, Vicente L., Mason, Robert A., Mitchell, Tom M., and Just, Marcel Adam. 2011. Commonality of neural representations of words and pictures. *NeuroImage*, **54**(3), 2418–25.
- Simanova, Irina, van Gerven, Marcel, Oostenveld, Robert, and Hagoort, Peter. 2010. Identifying object categories from event-related EEG: Toward decoding of conceptual representations. *PLoS one*, **5**(12), e14465.
- Simanova, Irina, Hagoort, Peter, Oostenveld, Robert, and Van Gerven, Marcel A. J. 2014. Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*, **24**(2), 426–434.
- Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. 2008. Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.
- Speer, N. K., Reynolds, J. R., Swallow, K. M., and Zacks, J. M. 2009. Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, **20**(8), 989–999.
- Sprouse, J., Schütze, C. T., and Almeida, D. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua*, **134**.
- Sudre, Gustavo, Pomerleau, Dean, Palatucci, Mark, Wehbe, Leila, Fyshe, Alona, Salmelin, Riitta, and Mitchell, Tom. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, **62**(1), 463–451.
- van Schijndel, M., Murphy, B., and Schuler, William. 2015. Evidence of syntactic working memory usage in MEG data. In: *Proceedings of the Sixth Workshop on Cognitive Modeling and Computational Linguistics*.
- Wehbe, Leila, Vaswani, Ashish, Knight, Kevin, and Mitchell, Tom. 2014a. Aligning context-based statistical models of language with brain activity during reading. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.
- Wehbe, Leila, Murphy, Brian, Talukdar, Partha, Fyshe, Alona, Ramdas, Aaditya, and Mitchell, Tom. 2014b. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS one*, **9**(11), e112575.
- Willems, Roel M. (ed). 2015. *Cognitive Neuroscience of Natural Language Use*. Cambridge, UK: Cambridge University Press.
- Yarkoni, Tal, Poldrack, Russell A., Van Essen, David C., and Wager, Tor D. 2010. Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences*, **14**(11), 489–496.

Zheng, Wei-Long, Zhu, Jia-Yi, Peng, Yong, and Lu, Bao-Liang. 2014. EEG-based emotion classification using deep belief networks. Pages 1–6 of: *IEEE International Conference on Multimedia and Expo*.

Zinszer, Benjamin D., Anderson, Andrew J., Kang, Olivia, and Wheatley, Thalia. 2015. How speakers of different languages share the same concept. Pages 2829–2834 of: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

4 Graph Theory Applied to Speech: Insights on Cognitive Deficit Diagnosis and Dream Research

Natália Bezerra Mota, Mauro Copelli, and Sidarta Ribeiro

Abstract

In the past ten years, graph theory has been widely employed in the study of natural and technological phenomena. The representation of the relationships among the units of a network allow for a quantitative analysis of its overall structure, beyond what can be understood by considering only a few units. Here we discuss the application of graph theory to psychiatric diagnosis of psychoses and dementias. The aim is to quantify the flow of thoughts of psychiatric patients, as expressed by verbal reports of dream or waking events. This flow of thoughts is hard to measure but is at the roots of psychiatry as well as psychoanalysis. To this end, speech graphs were initially designed with nodes representing lexemes and edges representing the temporal sequence between consecutive words, leading to directed multigraphs. In a subsequent study, individual words were considered as nodes and their temporal sequence as edges; this simplification allowed for the automatization of the process, effected by the free software *SpeechGraphs*. Using this approach, one can calculate local and global attributes that characterize the network structure, such as the total number of nodes and edges, the number of nodes present in the largest connected and the largest strongly connected components, measures of recurrence such as loops of 1, 2, and 3 nodes, parallel and repeated edges, and global measures such as the average degree, density, diameter, average shortest path, and clustering coefficient. Using these network attributes we were able to automatically sort schizophrenia and bipolar patients undergoing psychosis, and also to separate these psychotic patients from subjects without psychosis, with more than 90% sensitivity and specificity. In addition to the use of the method for strictly clinical purposes, we found that differences in the content of the verbal reports correspond to structural differences at the graph level. When reporting a dream, healthy subjects without psychosis and psychotic subjects with bipolar disorder produced more complex

graphs than when reporting waking activities of the previous day; this difference was not observed in psychotic subjects with schizophrenia, which produced equally poor reports irrespective of the content. As a consequence, graphs of dream reports were more efficient for the differential diagnosis of psychosis than graphs of daily reports. Based on these results we can conclude that graphs from dream reports are more informative about mental states, echoing the psychoanalytic notion that dreams are a privileged window into thought. Overall these results highlight the potential use of this graph-theoretical method as an auxiliary tool in the psychiatric clinic. We also describe an application of the method to characterize cognitive deficits in dementia. In this regards, the *SpeechGraph* tools were able to sensitize a neuropsychological test widely used to characterize semantic memory, the verbal fluency test. Subjects diagnosed with Alzheimer's dementia were compared to subjects diagnosed with moderate cognitive impairment, either with amnestic symptoms only or with damage in multiple domains. Also studied were elderly individuals with no signs of dementia. The subjects were asked to report as many names of different animals as they could remember within one minute. The sequence of animal names was represented as a word graph. We found that subjects with Alzheimer's dementia produced graphs with fewer words and elements (nodes and edges), higher density, more loops of three nodes, and smaller distances (diameter and average shortest path) than subjects in the other groups; a similar trend was observed for subjects with moderate cognitive impairment, in comparison to elderly adults without dementia. Furthermore, subjects with moderate cognitive impairment with amnestic deficits only produced graphs more similar to the elderly without dementia, while those with impairments in multiple domains produced graphs more similar to the graphs from individuals with Alzheimer's dementia. Importantly, also in this case it was possible to automatically classify the different diagnoses only using graph attributes. We conclude by discussing the implications of the results, as well as some questions that remain open and the ongoing research to answer them.

4.1 Introduction

Every day when we wake up, before talking with other people, we talk with ourselves using inner speech to remember what day it is, where we are, to make plans about what to do in the next few minutes or hours, who we are going to meet, or what we are supposed to do. When we recognize this “inner speech”

as coming from ourselves, we may simply call it “thinking.” However, sometimes this inner speech is not recognized as self, but rather as stimuli generated elsewhere; this is the basis of what we call psychosis. Sometimes past memories dominate this mental space, and we focus on past feelings of sadness, joy, fear, or anxiety. Past and future memories are mixed in these first moments even before any interaction with another person. This flow of memories and thoughts helps organize our actions and to soothe our anxiety and sadness, as we can plan future solutions to solve past problems. Organized, healthy mental activity allows old and new information to interact in order to support different actions that take experience into account in an integrated manner. But what happens with this flow of thoughts when we are unable to organize our inner space?

For centuries, psychiatry has described symptoms known as thought disorder that reflects disorganization of this flow of ideas, memories, and thoughts (Andreasen & Grove, 1986; Kaplan & Sadock, 2009). Those symptoms are related with psychosis, a syndrome characterized by hallucinations (when one perceives an object that does not exist; a sensorial perception without a real external object) and delusions (when one believes in realities that do not exist for other people; ideas or beliefs not real for their peers) (Kaplan & Sadock, 2009). There are many different causes for psychosis, such as the use of psychoactive substances or neurological conditions such as cerebral tumors or epilepsies. However, psychotic symptoms may occur without a clear cause, starting with a strange feeling or perception, getting worse, creating a confused reality hard to share even with the closest person, and causing major mental suffering.

In association with this strange reality, the patient can experience the feeling of fragmentation of thoughts, having difficulty to organize ideas or to follow a flow of memories, impacting the way to express what they are thinking or feeling, creating meaningless speech (symptoms known as “alogia,” and “poor speech”). This frequently reflects a mental disorder known as schizophrenia. In other cases, the person may experience another aberrant organization of thought, with higher speed of mental activity, associating different memories and ideas (known as “flight of thoughts”), creating a speech with large amount of words (a symptom known as “logorrhea”) that never reaches the main point. This pattern of thought disorder is common during the mania phase of bipolar disorder, a psychiatric condition mainly described by opposite mood cycles comprising depressive and manic phases. This speech pattern changes during depressive phases in the opposite direction (low speed of thought, fewer associations, fewer amount of words during speech). The speech content can reflect that strange psychotic reality on all those conditions with unlikely word association, but the organization of ideas reflected in the word trajectories reveals different directions of thought disorder, helping psychiatrists make differential

diagnosis between bipolar disorder and schizophrenia, predicting different life courses and cognitive impacts.

The description of these different patterns of thought organization perceived through language helped psychiatrists distinguish between two different pathological states and predict different life courses (with higher cognitive deficits for schizophrenia, first known as *Dementia Precox* [Bleuler, 1911]). However, recognizing these features subjectively requires a long-term professional training and adequate time with each patient to know each individual and avoid misjudgments. And even with the best evaluation conditions it is only possible to quantify those features subjectively, judging disease severity by grades on the psychometric scales such as BPRS and PANSS (Bech, Kastrup, & Rafaelsen, 1986; Kay, Fiszbein, & Opler, 1987). The differential diagnosis requires at least six months of observation during the first episode (First, Spitzer, Gibbon, & Williams, 1990), which means that the initial treatment may occur under considerable doubt regarding the diagnostic hypothesis. This lack of objective quantitative evaluation also negatively impacts the research strategies that aim to find biomarkers for complex psychiatric conditions (Insel, 2010).

Another condition that benefits from early diagnosis and correct interventions to prevent major cognitive damage is Alzheimer's Disease (AD) (Daviglus et al., 2010; Kaplan & Sadock, 2009; Riedel, 2014). Specific characterization of risk during preclinical AD requires specialized investigations and still challenges professionals in the field, due to a lack of a consensual description of each stage (Daviglus et al., 2010; Riedel, 2014). Failure to recognize AD early on can lead to a loss of opportunity to prevent cognitive decline (Daviglus et al., 2010; Riedel, 2014). In summary, the currently poor quantitative characterization of cognitive impairments related to pathological conditions such as psychosis or dementia hinders the early detection of these conditions. In this scenario, the new field of computational psychiatry has been proposing mathematical tools to better quantify behavior (Adams, Huys, & Roiser, 2015; Montague, Dolan, Friston, & Dayan, 2012; Wang, & Krystal, 2014).

To this end, natural language processing tools are particularly interesting. It is now possible to simulate the expert's subjective evaluation with better precision and reliability, either by quantifying specific content features such as semantic incoherence (Bedi et al., 2015; Cabana, Valle-Lisboa, Elvevag, & Mizraji, 2011; Elvevåg, Foltz, Weinberger, & Goldberg, 2007), or by analyzing the structural organization of word trajectories recorded from patients (Bertola et al., 2014; Mota et al., 2012; Mota et al., 2014).

4.2 Semantic Analysis for the Diagnosis of Psychosis

One useful tool used to characterize the incoherent speech characteristic of psychotic crises is called Latent Semantic Analysis (LSA) (Landauer & Dumais,

1997). The strange reality created during psychotic states impacts the coherence of the flow of words when patients express their thoughts freely, leading to improbable connections between semantically distant words within the same sentences.

LSA is based on a model that assumes that the meaning of each word is a function of its relationship with the other words in the lexicon (Landauer & Dumais, 1997). By this rationale, if two words are semantically similar, i.e., if their meanings are related, they must co-occur frequently in texts. It follows that if one has a large enough database of word co-occurrences in a large enough corpus of texts, it is possible to represent each word of that corpus as a vector in a semantic space, and their proximity in that space will be interpreted as semantic similarity (Landauer & Dumais, 1997).

When healthy subjects describe their normal reality, it is expected that they will use words that are semantically similar within the same text. However, when reality becomes bizarre, as typical of psychotic states, subjects are expected to use semantically distant words in sequence, thus building incoherent speech. That incoherence can be quantified as a measure of semantic distance between consecutive words or sets of words (for example, a set of words used in the same sentence). The more incoherent the speech, the larger the semantic distance between consecutive words or set of words. This was first shown for chronic patients with schizophrenia diagnosis (Elvevåg et al., 2007) and helped predict diagnosis in the prodrome phase, 2.5 years before the first psychotic crises (Bedi et al., 2015).

4.3 What Is a Speech Graph?

One way to quantify thought disorder is to represent the flow of ideas and memories reflected on the flow of words during a free speech as a trajectory and create a speech graph. A graph is a set of nodes linked by edges (formally defined as $G=(N, E)$, being $N=\{w_1, w_2, \dots, w_n\}$ and $E=\{(w_i, w_j)\}$ [Bollobas, 1998; Börner, Sanyal, & Vespignani, 2007]). The criteria determining how a link is established between two nodes define topological properties of these graphs that can be measured locally or globally. In the present case, each word is defined as a node, and the temporal sequence of words during a free speech is represented by directed edges (Mota et al., 2014) (Figure 4.1). From a speech graph we can objectively measure local and global features of the word trajectory that reflects the flow of thoughts during a free speech task (like when the subject reports a daily event, a past memory, or even a dream memory).

In the last decade, graph theory has been widely employed in the study of natural or technological phenomena (Boccaletti et al., 2006). By allowing the representation of the relationships among their units, the overall structure of a network can elucidate characteristics that could not be understood by

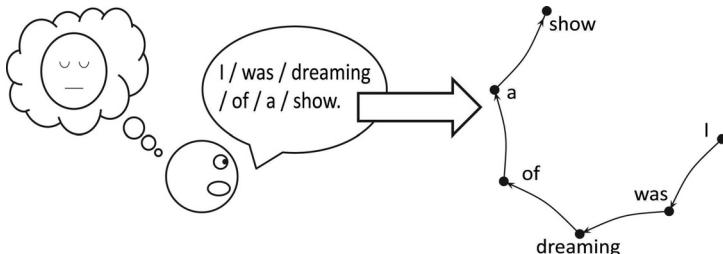


Figure 4.1 Examples of speech graphs from dream reports of schizophrenic, bipolar, and control subjects. Starting from transcribed verbal reports, graphs were generated using custom-made Java software (see the following text). Figure from Mota et al. (2014).

considering only a few units. The meaning of the represented structure basically depends on what is being considered as a node and on the definition of the presence and direction of edges (links between nodes). Graph theory as a tool not only may help tackle problems in the basic sciences but can also be applied to solve complex problems in everyday life, otherwise difficult to characterize and measure. An interesting strategy in scientific research is to keep both goals in focus: seek to understand a phenomenon at the fundamental level, while at the same time use the knowledge as a tool to solve practical problems (Stokes, 1997). With a simultaneous focus on basic and applied research, the application of graph theory to represent the relationship between spoken words helps understand how different psychiatric conditions differentially impact the flow of words during free speech, and how we can apply this knowledge to perform differential diagnosis.

Once reports are represented as graphs, one can calculate several attributes that quantify local and global characteristics. We calculated 14 attributes comprising 2 general graph attributes (Nodes and Edges), 5 recurrence attributes (Parallels – PE and Repeated Edges – RE; Loops of one – L1, two – L2 and three nodes – L3), 2 attributes of connectivity (Largest Connected Component – LCC and Largest Strongly Connected Component – LSC) and 5 global attributes (Average Total Degree – ATD, Density, Diameter, Average Shortest Path – ASP, Clustering Coefficient – CC) (Figure 4.2).

In order to compare graphs with different number of elements (controlling verbosity difference as measured by different amounts of words), two main strategies were used. First we divided each graph attribute by the amount of words in the report, assuming a linear relationship between graph attribute and verbosity. A pertinent critique is that the relationship between graph attributes and verbosity is not always linear, and for some attributes it is not clear if there is a direct relationship (Figure 4.3). A second strategy was to attribute a graph for

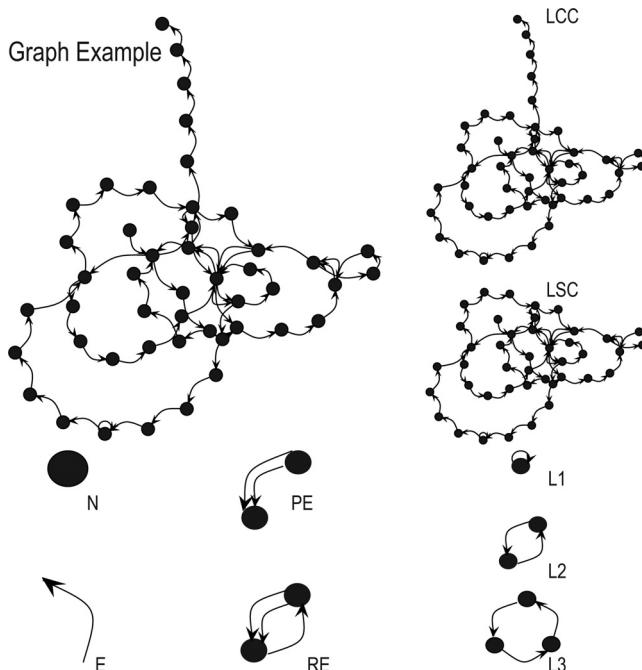


Figure 4.2 Examples of Speech Graph Attributes described earlier (figure from Mota et al., 2014).

Speech Graph Attributes:

1. **N:** Number of nodes.
2. **E:** Number of edges.
3. **RE (Repeated Edges):** sum of all edges linking the same pair of nodes.
4. **PE (Parallel Edges):** sum of all parallel edges linking the same pair of nodes given that the source node of an edge is the target node of the parallel edge.
5. **L1 (Loop of one node):** sum of all edges linking a node with itself, calculated as the trace of the adjacency matrix.
6. **L2 (Loop of two nodes):** sum of all loops containing two nodes, calculated by the trace of the squared adjacency matrix divided by two.
7. **L3 (Loop of three nodes):** sum of all loops containing three nodes (triangles), calculated by the trace of the cubed adjacency matrix divided by three.
8. **LCC (Largest Connected Component):** number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the underlying undirected subgraph. When you have all the words on one large connected component, LCC will be the same as N.

(continued)

each set of a fixed number of words, skipping an also fixed number of words to build the next graph, assuming a certain level of overlap between consecutive graphs. This “sliding window” approach allows calculating the average graph attributes of a graph with a fixed number of words. This enables the study of topological characteristics of graphs with different reports size (say, small, medium, and big graphs). A critique for this strategy is the arbitrary cut of word sequences that can change topological properties, mainly global attributes. This is an important discussion of ongoing research that needs to be addressed carefully, so as to enable a better interpretation of the results.

4.4 Speech Graphs as a Strategy to Quantify Symptoms on Psychosis

In an attempt to represent the flow of thoughts presented in a free speech, speech graphs were initially designed with nodes representing lexemes (a subject, object, or verb on the sentence), and their temporal sequence represented as directed edges, yielding directed multigraphs with self-loops and parallel edges (Mota et al., 2012). Analyzing dream reports represented as graphs from 24 subjects (8 subjects presenting psychotic symptoms with schizophrenia diagnosis, 8 subjects also with psychotic symptoms diagnosed as bipolar disorder in the mania phase and 8 control subjects without any psychotic symptom), it was possible to quantify psychiatric symptoms such as:

(Figure 4.2 caption continued)

9. **LSC (Largest Strongly Connected Component):** number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the directed subgraph (node a reaches node b, and b reaches a).
10. **ATD (Average Total Degree):** given a node n, its Total Degree is the sum of “in” and “out” edges. Average Total Degree is the sum of Total Degree of all nodes divided by the number of nodes.
11. **Density:** number of edges divided by possible edges ($D = 2*E/N*(N-1)$), where E is the number of edges and N is the number of nodes.
12. **Diameter:** length of the longest shortest path between the node pairs of a network.
13. **Average Shortest Path (ASP):** average length (number of steps along edges) of the shortest path between pairs of nodes of a network.
14. **CC (Average Clustering Coefficient):** given a node n, the Clustering Coefficient Map (CCMap) is the set of fractions of all n neighbors that are also neighbors of each other. Average CC is the sum of the Clustering Coefficients of all nodes in the CCMap divided by the number of elements in the CCMap.

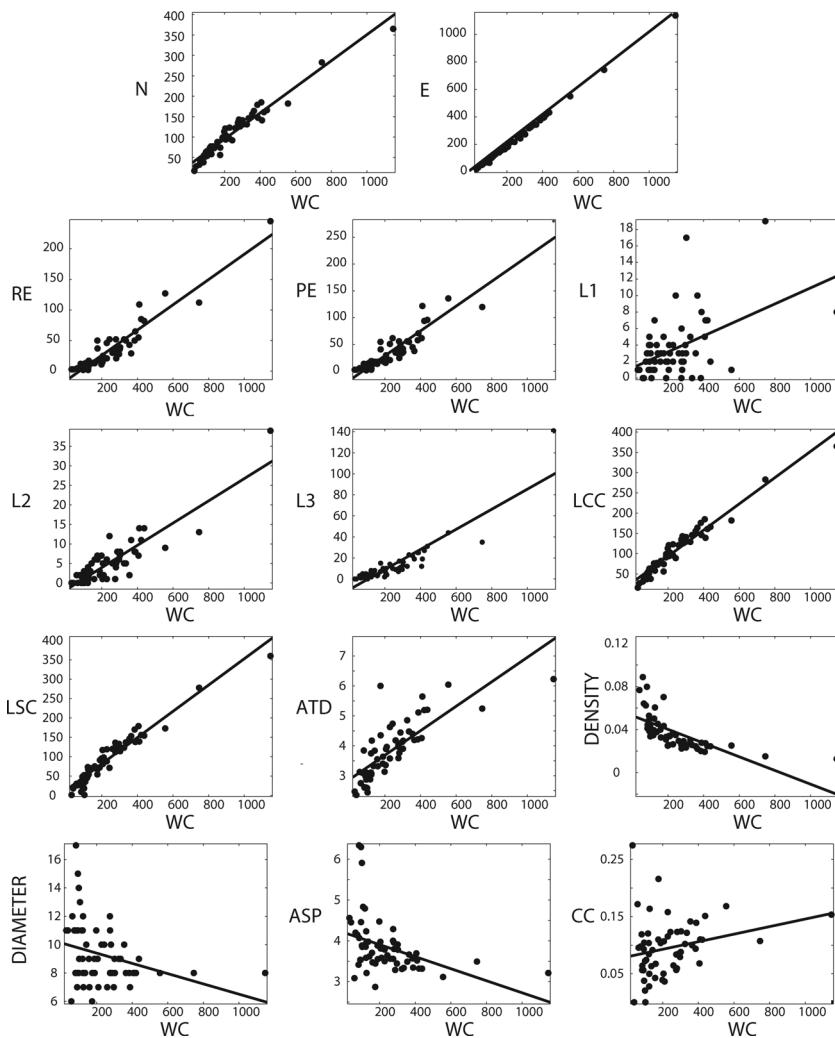


Figure 4.3 Linear correlation between SGA and word count (WC) (figure from Mota et al., 2014).

1. Logorrhea, described as the increase of verbosity characteristic of bipolar disorder on mania phase. This was quantified not only by counting more words in the bipolar group but also by more frequent recurrence (more parallel edges), even when controlling for differences in verbosity by dividing graph attributes by the amount of words in the speech. This means that the reports tend to return more often to the same topics.

2. Flight of thoughts, described as talking about other topics than the main topic asked, which is also characteristic of bipolar disorder. In the bipolar group, more nodes were used to talk about waking events upon request to report on a recent dream.
3. Poor speech, described as loss of meaning on the speech and perceived as a set of words that are poorly connected, characteristic of schizophrenia. This was quantified as more nodes per words, denoting reports that address the topics only once, neither branching nor recurring, so almost all the words used will count as a different node.

It was possible to automatically sort schizophrenia from bipolar group using a machine learning approach. A Naïve Bayes classifier was used to distinguish between both groups, and to distinguish between pathological groups and non-psychotic subjects (Kotsiantis, 2007). The classifier received as input either speech graph attributes or grades given from psychiatrists concerning psychiatric symptoms (using standard psychometric scales: PANSS [Kay et al., 1987] and BPRS [Bech et al., 1986]). Classification accuracy was assessed through the calculation of sensitivity, specificity, kappa statistics, and the area under the receiver operating characteristic curve (AUC), described as a plot of sensitivity (or true positive rate) on the y-axis versus false positive rate (or 1-specificity) on the x-axis. An AUC around 0.5 means a random classification, whereas AUC = 1 means a perfect classification (none of the possible errors were made). It was possible to classify the pathological groups against non-psychotic group using graph attributes and psychometric scales with high accuracy (AUC higher than 0.8) (Table 4.1). But to distinguish between schizophrenia and bipolar groups, graph attributes performed better than psychometric scales (AUC = 0.88 using graph attributes as input, while AUC = 0.57 when using psychometric scales as input) (Table 4.1).

This first study had some limitations concerning the low sample (only eight subjects per group) and the methodology. First, the transformation from a text to a graph was handmade, a process that is time consuming and has a higher risk of error. Second, the graph was not completely free of subject evaluation (a node

Table 4.1 Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2012).

	Sensitivity	Specificity	Kappa	AUC
S × B	93.8%	93.7%	0.88	0.88
S × C	87.5%	87.5%	0.75	0.90
B × C	68.8%	68.7%	0.37	0.80

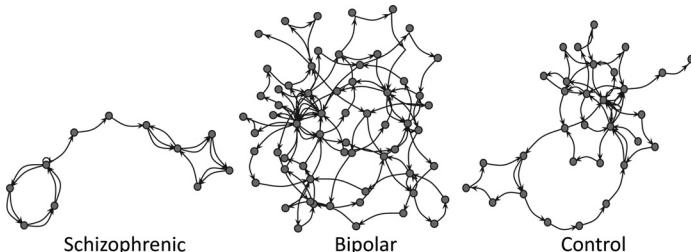


Figure 4.4 Representative speech graphs extracted from dream reports from a schizophrenic, a bipolar and a control subject (figure from Mota et al., 2014).

was considered as a subject, object, or verb on the sentence and, at a grammar level, it required a syntactic evaluation). So, in order to avoid these problems and to allow the study of a larger sample with larger texts, in a subsequent study we employed words as nodes and their temporal sequence as edges, a simplification that allowed the process to be automatized by the *SpeechGraphs* software (Mota et al., 2014). This custom-made Java software, available at <http://neuro.ufrn.br/softwares/speechgraphs>, receives as input a text file and returns the graph based on the text with all the 14 graph attributes described before. It is also possible to cut the text in consecutive graphs with a fixed number of words, controlling for verbosity and exploring different sizes of word windows to study cognitive phenomena.

To characterize distinct pathological phenomena in the speech of different types of psychosis, the *SpeechGraphs* tool was applied. Symptoms of Bipolar Disorder such as logorrhea could still be associated to the increase of the network size (Figure 4.4) (Mota et al., 2014; Mota et al., 2012). Also, symptoms of schizophrenia such as alogia and poor speech were measured as fewer edges (E), and smaller connected components (LCC) and strongly connected components (LSC) when compared to bipolar and control groups (Figure 4.4), producing less complex graphs in the schizophrenia group even after controlling for word count (comparing consecutive graphs of 10, 20, and 30 words with one word as step). In graphs from this group there are fewer edges between nodes and fewer nodes connected by some path or mutually reachable. This means that the schizophrenia group tends to talk only a few times about the same topic, not returning or associating past topics with consecutive ones, probably denoting cognitive deficits such as working memory deficits.

Using these network characteristics it was also possible to automatically sort the schizophrenia and bipolar groups, and those from subjects without psychosis, with $AUC = 0.94$ to classify schizophrenia and control groups, $AUC = 0.72$ to classify bipolar and control group, and $AUC = 0.77$ to classify schizophrenia and bipolar groups (Table 4.2). These results

Table 4.2 *Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2014).*

	AUC	Sensitivity	Specificity
S × B × C	0.77	0.62	0.81
S × B	0.77	0.69	0.68
S × C	0.94	0.85	0.85
B × C	0.72	0.74	0.75

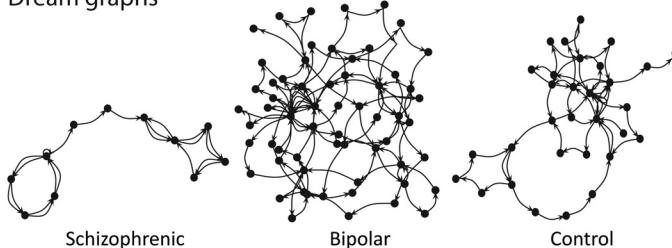
highlight the potential use of this method as an auxiliary tool in the psychiatric clinic.

To better understand the relationship between these graph features and the symptomatology measured by psychometric scales, the correlation between those metrics was analyzed. Edges, LCC, and LSC were strongly negatively correlated with cognitive and negative symptoms (as measured by psychometric scales). In other words, when the subjects presented more severity on symptoms such as emotional retraction and flattened affect (loss of emotional reaction), poor eye contact (with the interviewer during psychiatric evaluation), loss of spontaneity or fluency on speech, and difficulty in abstract thinking (measured by the ability to interpret proverbs), their reported dreams generated graphs with fewer edges and fewer nodes on the largest connected and strongly connected component. Those psychiatric symptoms are more common in subjects with schizophrenia (Kaplan & Sadock, 2009), indicating how we can measure the impact on cognition and deficits in social interactions of these individuals through graphs of speech (Mota et al., 2014). Cognitive and psychological aspects that drive this pattern of speech such, as working memory, planning, and theory of mind abilities, may explain those deficits and help elucidate the pathophysiology of the different psychotic disorders. When the interviewer asks the subject to report a memory, the way the subject interacts socially with the interviewer and recalls what to report, planning the answer and the sequence of events to report, impacts the sequence of words spoken, reflecting their mental organization.

4.5 Differences in Speech Graphs due to Content (waking × dream reports)

We already understand that during pathological cognitive states there is an impact on the flow of thoughts or memories that we can track by the word trajectory. But what happens with physiologically altered consciousness states like dream mentation? Is it possible to characterize differences between dream

Dream graphs



Waking graphs

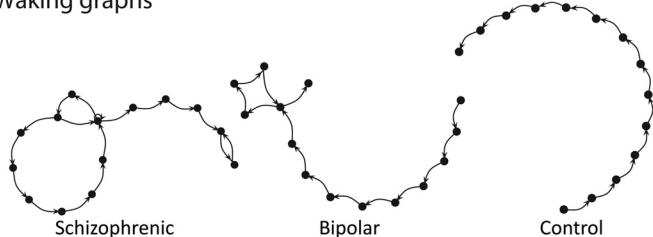


Figure 4.5 Representative speech graphs examples extracted from dream and waking reports from the same schizophrenic, bipolar, and control subject (figure from Mota et al., 2014).

and daily memories regarding word trajectories? Does it inform any additional features about general cognition?

A few minutes before waking up every day we can experiment an exclusively internal reality not shared with our friends or family: dreaming. This reality is internally built based on a set of memories with different affective valences, with different types of meaning only accessible by the dreamer. This confused mental state is phenomenologically similar to a psychotic state, as there is a lack of insight regarding the bizarreness of this strange reality (Dresler et al., 2015; Mota et al., 2014; Scarone et al., 2007). Thus it would not be surprising to expect that the flow of information regarding dream memories could better reveal thought disorganization characteristic of psychotic states.

During the studies with psychotic populations, there were differences in speech graphs depending on the speech content. When reporting a dream, subjects without psychosis and subjects with bipolar disorder produced more complex graphs (higher connectivity) than when reporting daily activities of the previous day, a difference that was not observed in subjects with schizophrenia (those subjects reported dreams or daily memories with the same few connected graphs) (Figure 4.5) (Mota et al., 2014). Therefore, graphs of dream reports were more efficient in group sorting than graphs of daily reports (Mota et al., 2014).

Another intriguing result was found in the correlations between speech graph attributes and clinical symptoms measured by psychometric scales PANSS (Kay et al., 1987) and BPRS (Bech et al., 1986). Only dream graphs connectivity attributes were strongly and negatively correlated with negative and cognitive symptoms (as measured by both scales) that are more common in schizophrenia. Waking report graphs showed negative correlations between general psychotic symptoms such as loss of insight (measured by PANSS) and incoherent speech (measured by BPRS) with LCC (also a connectivity attribute) (Mota et al., 2014). This emphasizes that reports of dream memories requires different cognitive functions and empathy abilities than reports of daily memories.

Based on these results we can conclude that graphs from dream reports are more informative about mental states than are graphs representing waking reports. This result echoes the psychoanalytic proposal that dreams are a privileged window into thought (Freud, 1900; Mota et al., 2014). This observation has started a new basic research approach to quantitatively understand what is going on when we remember a dream. The use of electrophysiological approaches (most notably, multichannel electroencephalography) to characterize different sleep stages in the laboratory allows the access to dream mentation by their reports at the same time that we access electrophysiological activity during sleep.

4.6 Speech Graphs Applied to Dementia

Considering the characterization of cognitive deficits in conditions such as dementia, the use of tests designed to characterize specific cognitive impacts on memory domain are useful in early evaluation. One example is the Verbal Fluency Test, which consists on verbal recall of different names of a specific category (usually animals) during a fixed time. This was first used to investigate the executive aspects of verbal recall, counting the capacity to produce an adequate quantity of words in a limited condition of recall, not repeating or recalling different categories (Lezak, Howieson, Bigler, & Tranel, 2012). The individual needs to access semantic memory correctly and to be flexible in order to quickly change the words (using temporal cortex structures), and to store the already mentioned words to avoid repetitions, which requires executive functions such as inhibitory control (using frontal cortex structures) (Henry & Crawford, 2004).

Different pathologies, such as dementia, can damage the performance on this task. As different structures are involved to correctly answer the task, different kinds of errors can help distinguish between different causes (damage in different locations). Different causes of dementia lead to different symptomatology

evolutions, which represent different location damages. The characterization of word trajectory with the application of the *SpeechGraph* tool complements this neuropsychological test (Bertola et al., 2014). A total of 100 individuals – 25 subjects diagnosed with Alzheimer's dementia, 50 diagnosed with Moderate Cognitive Impairment (25 of them with only amnestic symptoms and the others 25 with damage in multiple domains), and 25 elderly subjects with no signs of dementia – were asked to report as many names of different animals as they could remember in one minute (Nickles, 2001). The sequence of animal names was represented as a word graph.

It was observed that subjects with Alzheimer's dementia produced graphs with fewer words and elements (nodes and edges), higher density, more loops of three nodes and smaller distances (diameter and average shortest path) than did other groups, with the same trend for subjects with moderate cognitive impairment compared to elderly adults without dementia (Bertola et al., 2014). Furthermore, subjects with moderate cognitive impairment with only amnestic deficits produced graphs more similar to the elderly without dementia, while those with impairments in multiple domains produced graphs more similar to the graphs from individuals with Alzheimer's disease. Also in this case, it was possible to automatically classify the different diagnoses only from graph attributes (Bertola et al., 2014). There was also correlation between speech graph attributes and two important standard cognitive assessments widely used on geriatric population, denoting an important correlation between word trajectory on verbal fluency recall and general cognitive status (measured with MMSE – Mini Mental State Exam) and functional performance (measured with the Lawton Instrumental Activities of Daily Living Scale) (Bertola et al., 2014).

On one hand, the more cognitively preserved were the elderly, the more unique nodes were produced on less-dense graphs. On the other hand, the more functionally dependent the individuals were, the less words, nodes, and edges were produced on denser graphs with smaller diameter and average shortest paths (Bertola et al., 2014). Another differential impact was evident for three-node loops, a repetition of the same word with only two words in between (example: "lion," "cat," "dog," "lion"), found in higher frequency in the Alzheimer group compared with MCI and control groups (Bertola et al., 2014). This means an impairment in working memory since the early stages of the Alzheimer's disease (already recognized by other working memory assessments [Huntley & Howard, 2010]).

These results point to the additional information that the characterization of word trajectory brings to a well-established neuropsychological test. On this application example, as the test has restricted rules, we expect that the subject produces a certain type of graph, and different types of deviations from this expected pattern informs about cognitive impairments.

4.7 Future Perspectives

Word graphs are not the only tool to quantify psychiatric symptoms on speech analysis. As pointed out in the introduction, other approaches aim to quantify semantic similarities between words (Bedi et al., 2015; Elvevåg et al., 2007). The relationship between speech incoherence measured by LSA and speech structure measured by Speech Graphs is not clear yet. Both measures take into account word sequences and word co-occurrences, but with very different approaches (one compares with a semantic model based on a large corpus, and the other uses graph theory to characterize topological features of the speech sample). Understanding both approaches better can improve automated speech analysis for clinical purposes such as diagnosis and prognosis prediction, creating useful follow-up tools in a clinical set.

Other interesting perspective is to combine language analysis with prosody analysis. Semiautomated tools have characterized prosodic deficits related to schizophrenia diagnosis. The patients made more pauses, were slower, and showed less pitch variability and fewer variation in syllable timing, expressing a flat prosody when compared to matched controls (Martínez-Sánchez et al., 2015). The relationship between expressive prosody and language features during free speech can elucidate several cognitive characteristics subjectively perceived by well-trained psychiatrists (Berisha, Wang, LaCross, & Liss, 2015).

A better understanding of word trajectories in free speech can also be applied in settings other than the psychiatric clinic. As these tools show important correlations with cognitive deficits in psychosis and dementia, could it be useful to characterize cognitive development in a school setting? This kind of approach could help predict cognitive impairment early enough to allow quick intervention, preventing learning disabilities that later on would be harder to manage. This could also help quantitatively characterize cognitive development in a naturalistic manner.

Acknowledgments

The authors dedicate this chapter to the memory of Raimundo Furtado Neto, who made important contributions to the development of the SpeechGraphs software. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants Universal 480053/2013-8 and Research Productivity 306604/2012-4 and 310712/2014-9; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Projeto ACERTA; Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE); FAPESP Center for Neuromathematics (grant # 2013/07699-0, São Paulo Research Foundation FAPESP).

References

- Adams, R. A., Huys, Q. J., & Roiser, J. P. (2015). Computational psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*. doi:jnnp-2015-310737.
- Andreasen, N. C., & Grove, W. M. (1986). Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophrenia Bulletin*, 12(3), 348–359.
- Bech, P., Kastrup, M., & Rafaelsen, O. J. (1986). Mini-compendium of rating scales for states of anxiety depression mania schizophrenia with corresponding DSM-III syndromes. *Acta Psychiatrica Scandinavica Supplementum*, 326, 1–37.
- Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D., Copelli, M., & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *Nature Partner Journals Schizophrenia*, 1, 15030. doi:10.1038/npj schz.2015.30.
- Berisha, V., Wang, S., LaCross, A., Liss, J. (2015). Tracking discourse complexity preceding Alzheimer's disease diagnosis: A case study comparing the press conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*, 45, 3.
- Bertola, L., Mota, N. B., Copelli, M., Rivero, T., Diniz, B. S., Romano-Silva, M. A., Ribeiro, S., & Malloy-Diniz, L. F. (2014). Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Frontiers in Aging Neuroscience*, 6, 185. doi:10.3389/fnagi.2014.00185.
- Bleuler, E. (1911). *Dementia praecox or the group of schizophrenias*. (J. Zinkin, Trans.). New York: International Universities Press.
- Bollobas, B. (1998). *Modern graph theory*. Berlin: Springer-Verlag.
- Börner, K., Sanyal, S., & Vespignani, A. (2007). Network science. In B. Cronin (Ed.), *Information today* (pp. 537–607). Medford, NJ: ARIST.
- Cabana, A., Valle-Lisboa, J. C., Elvevag, B., & Mizraji, E. (2011). Detecting order-disorder transitions in discourse: Implications for schizophrenia. *Schizophrenia Research*. doi:S0920-9964(11)00233-7.
- Daviglus, M. L., Bell, C. C., Berrettini, W., Bowen, P. E., Connolly, E. S., Jr., Cox, N. J., Dunbar-Jacob, J. M., Granieri, E. C., Hunt, G., McGarry, K., Patel, D., Potokay, A. L., Sanders-Bush, E., Silberberg, D., & Trevisan, M. (2010). NIH state-of-the-science conference statement: Preventing Alzheimer's disease and cognitive decline. *NIH Consensus and State-of-the-Science Statements*, 27(4), 1–30.
- Dresler, M., Wehrle, R., Spoormaker, V. I., Steiger, A., Holsboer, F., Czisch, M., & Hobson, J. A. (2015). Neural correlates of insight in dreaming and psychosis. *Sleep Medicine Reviews*, 20, 92–99. doi:10.1016/j.smrv.2014.06.004.
- Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1–3), 304–316.
- First, M. H., Spitzer, R. L., Gibbon, M., & Williams, J. (1990). *Structured clinical interview for DSM-IV Axis I Disorders – Research Version, Patient Edition (SCID-I/P)*. New York: Biometrics Research, New York State Psychiatric Institute.
- Freud, S. (1900). *The interpretation of dreams*. (J. Strachey, Trans. and Ed.). London: Basic Books.

- Henry, J. D., & Crawford, J. R. (2004). A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology, 18*(2), 284–295.
- Huntley, J. D., & Howard, R. J. (2010). Working memory in early Alzheimer's disease: A neuropsychological review. *International Journal of Geriatric Psychiatry, 25*(2), 121–132.
- Insel, T. R. (2010). Rethinking schizophrenia. *Nature, 468*, 187–193.
- Kaplan, H. I., & Sadock, B. J. (2009). *Kaplan & Sadock's comprehensive textbook of psychiatry*. Baltimore, MD: Wolters Kluwer, Lippincott Williams & Wilkins.
- Kay, S. R., Fiszbein, A., & Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*(2), 261–276.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. In I. Maglogiannis, K. Karpouzis, M. Wallace, & J. Soldatos (Eds.), *Emerging artificial intelligence applications in computer engineering: Real world AI systems with applications* (pp. 3–24). Amsterdam: IOS Press.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.
- Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, C. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.
- Martínez-Sánchez, F., Muela-Martínez, J. A., Cortés-Soto, P., Meilán, J. J. G., Ferrández, J. A. V., Caparrós, A. E., & Valverde, I. M. P. (2015). Can the acoustic analysis of expressive prosody discriminate schizophrenia? *The Spanish Journal of Psychology, 18*(E86). doi:doi:10.1017/sjp.2015.85.
- Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16*(1), 72–80.
- Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., & Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Scientific Reports, 4*, 3691. doi:10.1038/srep03691.
- Mota, N. B., Vasconcelos, N. A., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One, 7*(4), e34928. doi:10.1371/journal.pone.0034928.
- Nickles, L. (2001). Spoken word production. In B. Rapp (Ed.), *What deficits reveal about the human mind/brain: A handbook of cognitive neuropsychology*: Philadelphia: Psychology Press.
- Riedel, W. J. (2014). Preventing cognitive decline in preclinical Alzheimer's disease. *Current Opinions in Pharmacology, 14*, 18–22.
- Scarone, S., Manzone, M. L., Gambini, O., Kantzas, I., Limosani, I., D'Agostino, A., & Hobson, J. A. (2007). The dream as a model for psychosis: An experimental approach using bizarreness as a cognitive marker. *Schizophrenia Bulletin, 34*(3), 515–522.
- Stokes, D. E. (Ed.) (1997). *Pasteur's quadrant – basic science and technological innovation*. Washington, DC: Brookings Institution Press.
- Wang, X. J., & Krystal, J. H. (2014). Computational psychiatry. *Neuron, 84*(3), 638–654.

Part III

Data Driven Models

5 Putting Linguistics Back into Computational Linguistics

Martin Kay

Abstract

Almost all practitioners of natural language processing make a crucial error that also besets much of Chomsky's argument about the poverty of the stimulus in first language learning, namely that we can discover all we need to know about language by examining sufficiently large quantities of it. The error is to ignore the crucial function of language in referring to things in the real and imaginary worlds. Speakers and hearers appeal to this in many ways. Translators rely on it to provide information that they must supply in the target text but which is only implicit in the source. Reference is one of several properties of language that relate parts of a text or a discourse that may not be adjacent to one another. To the extent that linguists are concerned with how people use language for communication, they are interested in processes in people's heads and, to the extent that they are concerned with processes, they are interested in computation. It is this, rather than engineering feats like machine translation, that gives computational linguistics its special role.

5.1 Explicit and Implicit Information

Alfred Charles William Harmsworth, 1st Viscount Northcliffe (1865–1922), was the owner of two British newspapers, the *Daily Mail* and the *Daily Mirror*. He is credited with first pointing out that “Dog Bites Man” is an unlikely headline, whereas “Man Bites Dog” might herald a newsworthy story. Modern journalists are referring to the same phenomenon when they point out that no one ever writes a story about a plane because it did not crash. The point is obvious to journalists and to ordinary citizens seeking to avoid newsworthy flights. But it is missed by the adherents of Noam Chomsky's linguistic theories and by practitioners of a branch of linguistic engineering that has come to be known as *natural language processing* (NLP). This is curious because, but for an interest in language, these two groups have little in common.

Chomsky argues against the behaviorist view of language acquisition espoused, for example, by Skinner, on the grounds that even very large quantities of raw linguistic data contain neither the kind nor the quantity of information that would enable a general learning device to acquire everything that it would have to know to use the system as humans do. A device that could learn a language would have to come to the task with much knowledge of how language works already built in. Babies are clearly not born knowing English or Chinese, but they are born with brains that are specially adapted to the kinds of structure that are found in English and Chinese. This is the argument from what is called the *poverty of the stimulus*.

Children learn language almost entirely from positive examples – examples used appropriately and effectively in normal communication. Many of the very reasonable, but incorrect, hypotheses that they might entertain about the way the language works could only be rejected on the basis of specific negative information. By way of illustration, he points out that there are two ways in which an English learner might suppose the sentence *The man who is hungry is ordering dinner* might be turned into a question. One is to move the second occurrence of the auxiliary verb *is* to the front, giving the expected result: *Is the man who is hungry ordering dinner?* But, if questions are formed by moving auxiliaries to the front, then why not move the earlier occurrence of *is* in this sentence? This gives *Is the man who hungry is ordering dinner?* But this is a mistake that children do not make. The claim is that the raw data does not contain the information that would be needed to reject it, so that some, at least, of it must be acquired from another source. We take note, in passing, of the unquestioned assumption underlying this example, that the two sentences – the assertion and the question – are related by an operation that transforms one into the other.

The argument from the poverty of the stimulus has had many and persuasive critics. The details are beyond the scope of this essay. More interesting to us is another argument about how children acquire language that receives very little attention, in part, perhaps, because it is so obvious. For all that children learn language easily, they would presumably not do so if the effort did not pay off in some fairly direct way. To be worth learning, language needs to be good for something. An unfortunate child that was confined to a dark room where it heard English sentences through a loudspeaker would presumably learn little English. A child learns to say *doggie* when it sees a dog, and is indeed corrected if it says *kitty cat* instead. It says *cookie* when it wants a cookie, sometimes with a definite payoff.

As Ferdinand de Saussure (1906–1910) famously pointed out, the link between a word and the things it can be used to refer to is completely arbitrary in all but a few very special cases. The Portuguese word *puxé* is pronounced substantially like the English word *push*, but it means *pull*. There would be no way to learn this from a loudspeaker in a dark room. The importance of

this for Chomsky's linguistics is considerable, but it is also not central to our present concerns. However, it is also of crucial importance for natural language processing, a great deal of which is based precisely on the idea that the knowledge required for many language-processing tasks, such as machine translation, can be learned entirely automatically from studying sufficiently large amounts of text. That most important function of language, the referential function, is ignored entirely.

There is, of course, an obvious and very serious objection to this line of argument, namely that the referential function, though clearly central to our everyday use of language, does not need to be provided for explicitly in purely linguistic activities such as translation. A French sentence that contains the word *chien* is probably talking about a dog. This function is filled in English by the word *dog*. So, translating the French sentence into English involves at least three entities: the French word, the English word, and the dog that connects them. However, crucial though the dog is to the understanding of the sentences, it is not crucial to the translation process. If the two words are connected through the animal, then they are connected, and we can do as second language learners routinely do and say simply the *dog* means the same thing as *chien*. We are, of course simplifying the story greatly. Words often have several meanings and thus several translations, and we must consider the context in order to decide which is in play in any particular place. The underlying argument, however, remains strong and must be taken seriously. Our claim, however, will be that it is false.

In what follows, we concentrate mainly on translation, but similar considerations apply equally to other kinds of language processing. The claim that the referential function of language does not need to be addressed explicitly in translation rests on the notion that translation is an essentially *syntactic* phenomenon. In other words, it consists in (1) deciding what lexical items the words and phrases of the source text correspond to, (2) finding words and phrases in the target language with corresponding lexical items, and (3) putting these words and phrases in an appropriate order. For these purposes, we can think of a lexical item as being essentially a pair consisting of a word or phrase in each of the languages. This model is implicit in almost all second language teaching, at least in the early years, and all work on machine translation, whether by practitioners of the new NLP or by linguists of the older rule-based tradition. It is in contrast with what we can refer to as *pragmatic* tradition, in which the translator adopts a position similar to that of the original author and, having understood what that author is trying to say, seeks a way of saying it in the target language, without attempting to relate individual words and phrases in one language with words and phrases in the other.

The work of most modern journeyman translators is largely syntactic, but problems requiring a pragmatic solution are quite common, and pragmatic solutions are often adopted even when syntactic solutions are readily at hand. When

mainline trains arrive at Paris, the loudspeakers generally address the passengers with the words: *Assurez-vous que vous n'avez rien oublié dans le train*, which can be translated syntactically into English as *Make sure that you have not forgotten anything on the train*. A marginally more idiomatic rendering would probably be *Make sure that you have not left anything on the train*, which is a slight concession to pragmatic translation, because the English verbs *forget* and *leave* would not normally be said to mean the same thing. They are equivalent in the present context because what they *refer* to is the same, namely articles that remain on the train even though the passengers intended to take them off with them.

As a matter of fact, the English message on the loudspeakers in Paris is often neither of these, but rather *Make sure that you take all your belongings with you*, which is clearly pragmatic. There is nothing in the original about belongings or about passengers having anything with them. If a passenger was carrying something for someone else, it is not clear that it should be counted among that passenger's belongings. A passenger could have left nothing on the train, but left it somewhere on the platform or had it stolen by a pickpocket, in which case he would not have it with him but would not have forgotten or left it on the train.

Professional translators generally strive for a result that is smooth and idiomatic so that it reads like an original document. They often feel that this is possible only if they resort to pragmatic translations. Sometimes it is quite unavoidable, most notably where the target language requires substantive information to be made explicit that is only implicit in the source text. These situations are so common, and the crucial pieces of information often seem so unimportant, that they are not recognized for what they are.

Consider a text that describes one person saying to another, *Is that your cousin?* We do not know exactly who is being referred to, but we take it that it is clear to the people involved in the conversation. If we have to translate this into French, we are faced with a serious problem, because the only words for *cousin* that French makes available are specific to one gender or the other. The sentences *Est-ce que c'est ton cousin?* and *Est-ce que c'est ta cousine?* are both good French, but they crucially contain information about the sex of the person being referred to. Someone might suggest *Est-ce que c'est ton cousin ou ta cousine?*, but this will not do because the people whose conversation is being reported do know who they are referring to. For the sake of the unfortunate translator, we must hope that the surrounding context contains information that resolves the issue.

Consider the somewhat simpler situation where the original is French. Let us suppose that the two people having the conversation are observing another pair – a man and a woman. One says to the other, *Est-ce que c'est ta cousine?* Since the word *cousine* can be used only for female referents, we know which

one of the two people is being referred to. If this is translated into English in the most straightforward way, we get *Is that your cousin?* in which the sex of the referent is no longer explicit. A translator who thinks this important must adopt a creative, pragmatic, solution such as *Is that woman your cousin?* unless the person is clearly a child, in which case it must presumably be *Is that girl your cousin?* Suppose the reply to the question is *Non, je n'ai pas de cousine.* This cannot be translated as *No, I do not have a cousin*, because the French speaker is denying only that they have a female cousin, leaving open the question of whether they have male cousins. We leave the translation of this sentence to the creative reader.

In the NLP framework, machine translation systems must be the result of an automatic process, referred to as *training*, applied to a corpus of existing translations that constitutes the *training data*. At present, this results in a *translation model* and a *language model*. The translation model consists essentially of the word and phrase pairs essential to syntactic translation, and the system learns these essentially by computing the probability of seeing particular words or phrases in a target sentence, given the presence of a particular word or phrase in the source. The assumption is that the greater the number of training sentences considered, the better the estimate of these probabilities will become. But this is clearly true only if the overwhelming preponderance of the target sentences are the result of syntactic translation. Every pragmatically translated sentence is a red herring, and the more of them that there are in the training data, the more the training process will be led astray.

An examination of the Europarl corpus, which has constituted that training data for countless statistical machine translation systems reveals that the majority of sentences consisting of more than a very few words contain at least some material that was translated pragmatically. The second sentence of the first text contains the words *I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period* opposite the French *je vous renouvelle tous mes voeux en espérant que vous avez passé de bonnes vacances*. There is nothing in the French that corresponds to *I would . . . like to* and we can only take it that the English *once again* is represented by the first two letters of the French *renouvelle*. The French *en espérant que vous avez passé de bonnes vacances* translated more or less word for word into English, would be *hoping that you have had a pleasant holiday*. Shall we say that *holiday* means the same as *festive period*?

The following sentence begins *Although, as you will have seen . . .* in English, and *Comme vous avez pu le constater . . .* (*As you have been able to ascertain . . .*) in French. The sentence after that begins *You have requested a debate . . .* which appears in French as *Vous avez souhaité un débat* (*You wanted a debate*). So it goes on. Remember that the systems that are trained using these data will operate in a strictly syntactic manner, as we have characterized it.

The Russian linguist, Roman Jakobson, one of the founders of the “Prague school” of linguistics, famously wrote: *Languages differ essentially in what they must convey and not in what they may convey*. Some require the sex of the person being referred to to be made explicit. The Bantu languages do not, but they do require information about its shape, which, in the case of abstract objects, for example, is just as conventional as gender is in Indo-European Languages. English nouns must be singular or plural, and verbs present or past. Definite or indefinite articles are required, with different consequences for the message conveyed. When looking for a translation of the common verb *go* into German, we must make explicit whether the going is on foot or in some sort of conveyance. Airplanes are distinguished for this purpose, from other kinds of conveyance. In French, one cannot talk of books without making it clear whether they are of the kind intended primarily to be read, to be written in, or if they contain tickets to be torn out. To translate *to be*, that most common of English verbs, into Spanish, one must determine whether the property being ascribed to the subject is temporary or permanent.

Practitioners of NLP, as that term is generally understood, agree that all the information required for tasks like translation is implicit in the translations themselves so that a computer program can be made to learn how to translate from examples previously produced by human translators. Some of the information may be in very weak dilution so that considerable ingenuity and massive amounts of data may be necessary to recover it. Counterexamples are sufficiently rare to be negligible. The examples we have adduced, however, involve some of the commonest lexical and grammatical phenomena, and their resolution seems to require real experience of the real world by the translator. A rich source of further examples is to be found in what Hector Levesque has called *Winograd schemata* after the Stanford computer scientist Terry Winograd.¹ These consist of pairs of sentences, of which the following is the most well known:

The city councilmen refused the demonstrators a permit because they feared violence.

The city councilmen refused the demonstrators a permit because they advocated violence.

The question is: Who does the pronoun *they* refer to in each sentence – the councilmen or the demonstrators? Subjects agree overwhelmingly that it is the councilmen that fear, but the demonstrators that advocate. If one of these sentences is to be translated into a language in which the pronoun is required to manifest some kind of agreement, say in gender, with its antecedent, and if the

¹ Terry Winograd, *Understanding Natural Language*, Academic Press, 1972. See also Gary Marcus: “Why can’t my computer understand me?” *The New Yorker*, August 23, 2013.

two potential referents differ in this property, then the translator must solve the riddle. In this case, what is presumably required is general knowledge about demonstrations, councilmen, and violence. In other words, successful translation is possible only by a person or a mechanism that, in a serious sense, understands the texts it works on.

There is a striking contrast between what we can learn from newspaper headlines about dogs and men biting one another on the one hand and Winograd schemata on the other. The headline conveys information about something that is, at least to some extent, surprising. The information is laid out as clearly and explicitly as possible. In the Winograd schemata, the information we are concerned with is secondary, unsurprising, and inexplicit. Not surprisingly, they present quite different problems to the designer of language-processing systems. A well-known statistical machine translation system was given the English sentence *He had eaten earlier that evening* and invited to translate it into Spanish. It delivered the result *No había comido esa misma tarde*, which means, of course, exactly the opposite. Presumably this is because a person's failure to eat is generally more newsworthy, and therefore more frequently remarked upon, than their eating, which people are generally expected to do regularly. The word *no* is introduced because the sequence *había comido* has been seen rarely, if ever, without it. The inescapable conclusions are that statistically based systems cannot extract information that is not explicit in the source text because it simply is not there to be extracted from data that has no referential component, and it will often treat explicit information incorrectly because it is precisely the divergence of a text from what is expected that makes it worth writing in the first place.

Consider the schema:

They could not get the book in the box because it was too big.

They could not get the book in the box because it was too small.

What was to big or two small? The book or the box? Most of us have had enough experience trying to put objects into containers to know how their relative sizes can influence the expectation of success. The problem does not have to have a particular grammatical form that immediately labels it as a Winograd schema. The sentences could have been: *Nobody thought there would be any problem packing the computer up, but, when they tried to get it into the box, ...*

they found that it was too big.

they found that it was too small.

The sentences could be translated into French somewhat as follows: *Personne ne croyait que l'on auraient des problèmes à emballer l'ordinateur, mais quant ils essayaient de le mettre dans la boite ...*

il s'est révélé trop grand.
elle s'est révélée trop petite.

Let us be clear about the nature of this argument. The claim is not that, no matter how much text a machine were to read, it would never encounter enough accounts of attempts to put objects into containers, with explicit information on how the outcome was affected by their relative sizes, to support the necessary inference. Let us suppose that the training data was there, and in sufficient quantity. The claim we are making is that, without explicit experience with using boxes as containers, successfully and unsuccessfully, one would have no way of even recognizing these accounts as relevant. More particularly, there would be no way of recognizing them as important for gender agreement in a French translation.

The force of Chomsky's argument from the poverty of the stimulus in human language acquisition is complex and doubtless will be discussed for a long time. What seems altogether less problematic is its relevance to natural language processing. The model of language that aspiring members of the NLP community must embrace is indistinguishable from that of the child listening to language coming through a loudspeaker in a dark room. There must be no chink in the wall that would allow some light to fall on the referential component that gives the system its entire value. The information thus abjured cannot be recovered by amassing greater amounts of text or bringing greater ingenuity to its processing. It simply is not there.

5.2 Features

Since the fourth century BC, when Pāṇini wrote his Sanskrit grammar, linguists have been at pains to locate the components of language – sounds, words and parts of words, phrases, and sentences – in a space, and thus to give substance to the appearance of similarity and difference among them, and to the observation that similar components tend to behave in similar ways.

Vowels are high, mid, or low. Independently, they are front or back and, on a third dimension, they are rounded or unrounded. The similarities in what the articulators must do in order to produce vowels correspond closely to those among their acoustic properties and the functions they fill in various languages. Consonants arrange themselves on another set of dimensions. The initial consonants in the words *pot*, *tot*, and *cot* are *voiceless obstruents*, or *unvoiced stops*. They are all produced by briefly interrupting the flow of air through the mouth and then releasing it suddenly. They are voiceless because, unlike their voiced counterparts in *bot*, *dot*, and *got*, they are pronounced without moving the vocal cords. In English, there is release of air following the consonant when an unvoiced obstruent begins a word or a stressed syllable, except when

the obstruent is preceded by an *s* as in *spade*, *stake*, and *skate*. These are generalizations that are easy to make and to state given a prior classification of consonants in which unvoiced obstruents constitute a natural category. They also make it natural to observe and to describe other languages in which the release of air, for example, is absent.

Linguists recognize multidimensional feature spaces on every level of linguistic analysis. Indo-European morphology has a number dimension with singular, plural, and sometimes dual, a case dimension with nominative, accusative, and up to six others in different languages, and a gender dimension with masculine, feminine, and neuter. In syntax, there are declarative and interrogative sentences, active and passive sentences, noun phrases and verb phrases, and so forth. A fundamental part of linguistics is thus a classification of linguistic phenomena that applies to all languages and that provides a set of *features* in terms of which to describe these phenomena in individual languages.

If a natural language-processing system can be constructed on the basis of a system of features, then one with the same capabilities can, in principle, be built without them. All occurrences of a term like *accusative noun* in the design of the first system would simply be replaced by a list of all accusative nouns in the design of the second. A system built in this way would embody none of the generalizations and insights that linguists see as central to their field but, if the object of the enterprise is to engineer a system to fill a practical need, this is a secondary issue at best.

However, that this might be done in principle by no means implies that it is what should be done in practice. At some point during the training of a system intended to translate between English and German, it might discover that the English word *dog* can be put together with the German word *Hund* to form a lexical unit. In a quite unrelated event, it might conclude that another lexical unit can be formed with *dogs* and *Hunde*. It might learn that *dogs* could also be paired with *Hunden* because, as we are quietly reminded by the linguist inside us, *Hunde* is the plural of *Hund* except in the dative case, when it is *Hunden*. A linguist would see the events in these three classes as different from one another, but he would see their similarities as more important than their differences. In particular, he would doubtless be led to postulate a different kind of lexical unit relating not two different English words and three different German words, but a single English lemma and a single German lemma.

A lemma belongs to a family of words, *dog* and *dogs* in English, and *Hund*, *Hunde*, and *Hunden* in German. The number of lemmas in a text is generally smaller than that of words because there are less of them. The amount of information that each provides is therefore greater. The concomitant advantage is greater for morphologically rich languages than morphologically poorer ones – greater for German than for English, greater for Finnish than for German, and greater for Turkish than for Finnish.

In the early stages of studying German, students learn that the objects of certain prepositions, such as *von*, are always in the dative case, and they can apply this rule to all the nouns they encounter, even if they have never actually seen them in the dative plural. Students of German learn very early how to recognize and construct the dative plural forms of regular nouns. They also learn that the plural forms for the other cases are all the same. Also, in the singular, most nouns have at most two forms. In this respect, the linguistically informed human language learner lives in an incomparably richer and more secure environment than the machine for whom different spellings imply different words.

Zipf's law is a mathematical codification of the observation that very few words and phrases occur very frequently in texts, and a large number occur extremely rarely. Indeed, as the sizes of the texts one considers increases, the number of words that occur only once increases also. If we think of a person's entire experience with a language as a single text, this means that there will also be a large number of words that that person has seen in only one of its possible forms. In a frequently cited paper, Eugene Charniak² describes automatically extracting rules of a context-free grammar from a corpus of 300,000 words. The grammar contained 10,605 rules, of which 3,943 occurred only once. Rules are subject to Zipf's law just as words are. As Charniak says, *At about eleven thousand rules, our grammar is rather large* and, of course, correspondingly uninformative.

The observations we are making here support the view that the generalizations linguists are at pains to formalize in the course of analyzing a language are also made by speakers in the course of learning the language. We have argued that there is great practical advantage in generalizations about vocabulary items and their morphological variants. But there is altogether greater advantage that comes with generalizations and features at the higher levels of linguistic organization. Mastery of sentence structure, for example, whether for the linguist or the language learner, rests crucially on grasping the notion of locality that is appropriate to the words in a sentence. This notion of locality is relative not to proximity in the string of words, but to proximity in a recursive structure which gives the sentence its coherence. Consider the following sentences:

1. She added all the things she had bought up the same street.
2. She added all the things she had bought up the same way.
3. She added all the things until she had bought up the whole street.

The first sentence claims that she added things and that the things she added were bought *up the same street*. The words *up the same street* constitute a single entity in the recursive structure of the sentence, as do the words *the same street*.

² Eugene Charniak. 1996. "Tree-bank Grammars" Technical Report CS-96-02, Department of Computer Science, Brown University.

So the word *up* is very close to *the same street* in this structure. In sentence 2, on the other hand, the word *up* is the second part of a two-part lexical item of which *added* is the first part. These two words are therefore more closely bound to one another than either of them is to the words *all the things she had bought* that separate them in the string. In sentence 3, *up* is also the second part of a complex lexical item. Here, however, the first part is *bought*. Up is closer to *bought* than it is to *the whole street*.

Together with the aforementioned three sentences, consider the following pair:

4. She put any member of the family that came down up for the night.
5. She put any member of the family that came up down as a scoundrel.

Sentence 4 contains the lexical items *put up* and *come down*. Sentence 5 contains *come up* and *put down*.

Generalizations like these go back at least as far as Pāṇini and their importance for ordinary language learners as well as linguistic theoreticians was taken for granted from the fourth century BC until the advent of natural language processing at the end of the 20th century. The machine translation systems developed within this framework place much of the burden on a component known as the *language model* that is charged with selecting the most promising of the candidate translations proposed by the so-called *translation model*. Language models are, according to the commonest conception, all about the proximity of words in the string. Roughly speaking, one candidate will be considered better than another if it contains more and longer substrings that were also found in the training texts. The more often they were found there, the better. On such criteria, the sequences *came down* and *came up* might well be recognized as essentially idiomatic, but it is hard to see how *put . . . up* and *put . . . down* might be so recognized.

As translations of the aforementioned sentences, Google Translate gives the following in French:

6. Elle a mis tout membre de la famille qui est descendu pour la nuit. (*She put each member of the family that descended for the night*)
7. Elle a mis tout membre de la famille qui est venu vers le bas comme un scélérat. (*She put each member of the family who came downwards like a scoundrel*)

As German translations, Google gives:

10. Sie legte ein Mitglied der Familie, die für die Nacht kam. (*She laid a member of the family who came for the night*)
11. Sie legte ein Mitglied der Familie, die nach oben nach unten als Schurke kam. (*She laid a member of the family who came from above to below as a scoundrel*)

The information in the recursive structure of sentences can easily conflict with what the string, considered simply as a linear sequence, seems to imply. The sentence *The man with the dog fought for his life* contains the sequences *the dog fought* and *fought for his life*. The second is aligned with the true structure of the sentence, but the first is not. However, it is reasonable to suppose that both sequences would be well represented in a training corpus of interesting size. A sentence that is very similar to this, but which lacks the ambiguity, is *The man with the dog fought for her life*. We still do not know whose life is being fought for – the life of the dog or that of some previously mentioned person or animal – but we do know that it was not the man. The ambiguity would have to be resolved if the sentence were to be translated into Russian where in the sentence *He fought for his life*, the pronoun used to translate *his* would be different depending on whether it referred to the same person as *he*.

One may counter this argument by pointing out that natural language processing is in its early stages. Many of its proponents would readily acknowledge the importance of syntax and of recursive structure. It is just that we have not yet discovered reliable ways of learning how to recognize that structure from raw data. If this situation persists, we will go back to the linguists, some of whom are among our best friends, and have them annotate texts so as to make their structure explicit, and will have our systems acquire this part of the knowledge they need from the resulting annotated corpora. By having them annotate texts rather than coming right out and explaining to us how it is done is that we will still be able to claim that the eventual system will be the product of machine learning, a position to which we are religiously committed.

The disadvantage of annotation is that it is subject to Zipf's law so that each annotation that is made contains less information than the one before it. But the approach cannot be dismissed casually. The best part-of-speech taggers, after all, make hardly any more errors on a given task than experienced humans, and they were learned from texts annotated by human taggers. Furthermore, the arguments we have been making about the importance of recursive structure for recognizing lexical units, and for translation in general, presumably apply equally to the task of part-of-speech assignment. So let us see how the Stanford tagger, widely acknowledged to represent that state of the art, behaves. For the input *She put any member of the family that came down up for the night*, it returns *She_PRP put_VBD any_DT member_NN of_IN the_DT family_NN that_WDT came_VBD down_RP up_RP for_IN the_DT night_NN*. Our main interest is in the words *up* and *down*, which are both correctly tagged as particles (RP). The same goes for *She put any member of the family that came up down as a scoundrel* which is tagged *She_PRP put_VBD any_DT member_NN of_IN the_DT family_NN that_WDT came_VBD up_RP down_RP as_IN a_DT*

scoundrel_NN. Once again, *up* and *down* are tagged as particles. Unfortunately, to say that something is a particle is to say no more than that is a one member of a pair. Without the other member of the pair, this information can engender nothing but frustration.

The levels of abstraction on which linguists study language make contact with the outside world on the lowest level where they face sounds, articulatory gestures, and character shapes, and at the top level, where they encounter references to concrete and abstract objects in the outside world. It is at the top level that the child in the dark room and the machine that sees only text suffer most spectacularly from their abandonment by science and understanding. Consider the following scenario, which is not fictional. A lady comes into a railroad car on a train that is about to leave Montpelier and asks one of the passengers, *Does this train go to Perpignon?* The passenger replies, *No, it stops in Béziers*. Two obvious alternatives would suggest themselves to someone who, against all odds, were in the position of translating this into German, namely *Nein, er hält in Béziers* and *Nein, er endet in Béziers*. They are incompatible because *hält* implies that the train stops briefly in Béziers, and then resumes its journey, possibly toward Perpignon, while *endet* implies that Béziers is the train's terminus. A translator that was familiar with the rail system in southwestern France would know that Béziers is indeed on the line from Montpelier to Perpignon and that, while that is the terminus for some trains, others continue on to Perpignon. What is in question is therefore almost certainly that Béziers is the train's terminus so that *hält* is the best choice for the German verb.

The alternative approach to the problem has the advantage that it would be open to translators without knowledge of the local geography or train schedules. Notice that the single word *No* would have been an entirely adequate answer to the lady's question. The rules of polite discourse, however, require some explanation to be given, and the remark about Béziers fills this function. If the train made a brief stop in Béziers before continuing toward Perpignon, then the reply would not constitute an explanation, so that a German translation using the word *hält* would make no sense.

The second of these translation strategies illustrates what has been referred to as the cooperative principle in language use. Language could presumably not be used with carefree abandon in everyday situations if each utterance had to be precisely crafted like a mathematical formula to fulfill its purpose. People who want to understand must therefore be understanding. They must construe the utterances they hear in the ways in which they think the speaker is most likely to have intended them. A good listener is therefore someone who brings to the task as much knowledge as possible of the subject matter in general, the potential referents of the words and phrases, and good judgment about the interlocutor.

5.3 Linguistic Computation and Computational Linguistics

We have been trying to promote the view that linguistic computation should be distinguished from computational linguistics. Proponents of natural language processing hold to the term *computational linguistics* presumably because it has provided them with a most effective Trojan horse in which to penetrate linguistics departments in universities as well as professional societies and conferences. Those involved, however, apparently either dismiss attempts to understand how language actually works as irrelevant or persuade themselves that, by achieving high scores in public competitions, they actually transform statistical ingenuity into linguistic insight. Some members of the fraternity were competent linguists in an earlier life. They may perhaps be clinging to the forlorn hope that these two almost unrelated enterprises may one day come together, their discoveries fusing and taking us to a level of understanding beyond anything we can now contemplate.

For the most part, proponents of natural language processing see themselves as engineers, and are happy to be judged by what they contribute to the solution of practical problems. They have given us Google Translate, many and varied information retrieval systems, and programs that can tell us if a consumer review was favorable to the product or not. They do not begrudge old-style linguists in general, and computational linguists in particular, their interest in what they call *science*. But it behooves the linguists to say how it is that computation plays such a central role in their endeavors. After all, physicists, psychologists, and accountants routinely use computers, and special techniques and procedures have been developed to fill their requirements. But, as has often been pointed out, we do not distinguish computational physicists, psychologists, and accountants. Why, then, do computational linguists cling to this term?

The historical origins of the term *computational linguists* may make its continued use today look incidental and perhaps a little cynical. People who were working on machine translation when the ALPAC report³ appeared in 1966 foresaw the imminent disappearance of their livelihood if they could not rapidly recast themselves as the scientists that the report said would be required to give machine translation the foundation it desperately needed. If they were to be the practitioners of a new science, the most important thing that that science would need was a name. At least three names were proposed: *Automatic Natural Language Data Processing*, *Computational Linguistics*, and *Mechanolinguistics*. In the interest of good public order, we withhold the identities of the advocates for each of these proposals.

³ John R. Pierce, John B. Carroll, et al., *Language and Machines – Computers in Translation and Linguistics*. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

To say that we have come to be known as computational linguists by accident is not to say that the name is inappropriate or unwarranted. A language is not just a corpus. It is not simply a set of rules or organizational principles. In addition to these things, which characterize what is traditionally known as its *paradigmatic* component of language, it is a set of processes that take place in people's heads each time they produce or understand an utterance or write or read a sentence. These constitute the *syntagmatic* component. The paradigmatic component is at the service of the syntagmatic component giving, as it does, the structure shared among speakers of a language that enables them to understand one another.

There is a complex interplay between the paradigmatic and the syntagmatic components of language. The more permissive the rules and organizational principles, the more complex the syntagmatic processes, and the less psychologically plausible the resulting overall model. The study of processes, as abstract entities, is called *computer science*. The question of how a set of rules and organizational principles might be brought to bear on the problem of producing an utterance or understanding a sentence is in large measure one of algorithm design and computational complexity. We should by no means be taken as claiming that human language processors are essentially von Neuman computers or Turing machines. Computer science is about machines that exist and machines that are possible. In short, it is about the notion of process in a pure and abstract form, and this is at the very core of linguistics as it should and must be.

The approach to linguistics proposed by Zelig Harris (1900–1992) and developed by Chomsky and his followers is all about process. Sentences, for example, are described and explained in terms of the process that is assumed to result in their construction. Families of sentences, related typically by similarities in their meaning, start with a common form, which is then modified in a series of steps in which words and phrases are moved from one place to another in the structure. Movement is at the core of the system. An entity that would normally be expected to move from one place to another in the structure may be prevented from doing so because the location to which it would move is already occupied by another entity. The exact order of events is crucial, and everything turns on finding rules or, preferably, more general constraints on the sequences of events so that only the observed outcomes are possible.

The proponents of these systems usually refer to them as *generative grammar*, though they do not have an exclusive claim to this term. Systems that are similar to them in broad outline are also familiar to computer scientists. A *compiler*, for example, translates programs from the language in which the programmer writes to a minutely specified set of operations that the computer must carry out in exactly the order specified to compute the result the programmer desires. Programmers do not specify this sequence of events directly because the level of detail required is unnecessarily burdensome and obscures the

overall intent of the program. The language in which the programmer writes allows the sequence of events to be specified in much more general terms.

The models of language envisaged by Chomsky and his followers are similar to computers in requiring steps to be specified in minute detail. They differ to the extent that computer scientists recognize that there is a point beyond which detail obscures more than it reveals. Mathematicians do not attempt to explain the notion of quotients by describing the process of long division. Indeed, the only way a student can reasonably hope to understand long division is to first get a firm grasp of quotients. The possibility of bypassing long division that computers and calculators provided was seized immediately, and now we hear of it no more.

Computational linguistics has been, in large measure, an attempt to bring to linguistics some of the advantages that computers allowed in the teaching of arithmetic and that compilers brought to computer science. Given that the field is barely 60 years old, its successes have been remarkable. It has been successful in providing a number of models of linguistic processing whose computational power matches what the task seems to require extremely closely. It is powerful enough to allow all the necessary computation to be specified but constrained enough to capture important insights into the fundamental nature of linguistic processes.

Computational linguistics has given us a version of finite-state technology that is apparently a remarkably good fit for the requirements of morphology and morpho-phonology. It does not require the linguist who uses it to think in terms of states and transitions, but rather in terms of a sequence of representations of a word, each related in a straightforward way to the ones immediately preceding and following in the sequence, connecting an abstract representation at one end of the sequence with the observed representation at the other. This has become a common way of thinking among formal linguists, who generally find it very congenial.

More striking are the contributions that computational linguists have made in syntax, where such formalisms as Lexical Functional Grammar, Head Driven Phrase Structure Grammar, and Combinatory Categorial Grammar are widely acknowledged as major contributions. In these frameworks, Chomsky's problem of how to form a question from the sentence *The man who is hungry is ordering dinner* simply does not arise. The assertion and the question are indeed assumed to have similar underlying structures, but neither has a privileged position relative to those structures.

5.4 Conclusion

Nontrivial tasks, like translation, are AI-complete. In other words, any machine that can perform them would have to be able to mimic all aspects of human

intelligence. This is nowhere more evident than in machine translation. At no time since the launch of Sputnik caused serious work to begin on this problem has it been thought important that workers in this field should inform themselves about what a translation is generally taken to be by those who use them or what it is that distinguishes a professional translator from someone who took some French at school. It was, however, thought to be useful to know something of what linguists have discovered about language. Not only is any such knowledge now thought to be unnecessary; it is often regarded as a hindrance. With such knowledge, we will reject the possibility that *He had eaten earlier that evening* might, on rare occasions, be rendered into Spanish as *No había comido esa misma tarde*, and certainly that *man bites dog* is a possible, if rare, equivalent of *dog bites man*.

As we have already noted, one of the most remarkable properties of human language is that, despite its great subtlety and complexity, it is a tool that humans can use in an entirely casual manner. Communication in this medium is a collaborative enterprise in which the receiver is constantly second-guessing the sender's intentions. It is, therefore, a fundamentally probabilistic enterprise. When we say that nontrivial linguistic tasks are AI-complete, we are also saying that they are probabilistic. During most of the time since the launch of Sputnik, probability and statistics have been largely absent from linguistics. This is unfortunate and it doubtless did much to hamper progress in the field. But, slowly and cautiously, it is a matter that is being set to rights. There are now rooms in university linguistics departments with signs on the door saying things like *Syntax Laboratory*. Linguists use computer tools to search large corpora for examples, and the judgments of individual informants are no longer taken as sacred truth. But linguists have not embraced big data unquestioningly or replaced thought and analysis with machine learning. We must hope they never do.

6 A Distributional Model of Verb-Specific Semantic Roles Inferences

Gianluca E. Lebani and Alessandro Lenci

Abstract

In a standard view, commonly adopted in psycholinguistics and computational linguistics, thematic roles are approached as primitive entities able to represent the roles played by the arguments of a predicate. In theoretical linguistics, however, the inability to reach a consensus on a primitive set of semantic roles led to the proposal of new approaches in which thematic roles are better described as a bundle of more primitive entities (e.g., Dowty, 1991; Van Valin, 1999) or as structural configurations (e.g., Jackendoff, 1987). In a complementary way, psycholinguistic evidence supports the idea that thematic roles and nominal concepts are represented in similar ways (McRae et al., 1997b; Ferretti et al., 2001), thus suggesting that the former can be accounted for as predicate-specific bundles of inferences activated by the semantics of the verb (e.g., the patient of *kill* is typically alive before the event and dead afterward). Such inferences can take the form of either presuppositions or entailment relations activated when a filler saturates a specific argument position for a given predicate.

Our aim in this chapter is twofold. First, we report behavioral data collected to obtain a more fine-grained characterization of the thematic role properties activated by a subset of English verbs. To this end, we employed the modified version of the McRae et al. (1997b) elicitation paradigm proposed by Lebani et al. (2015) to describe which semantic properties of the participants are more relevant in each phase of the action described by the predicate. Next, we test the possibility to model such verb-specific inference patterns by exploiting corpus-based distributional data, thus proposing a novel approach to represent the same level of semantic knowledge that is currently described by means of a finite set of thematic roles.

6.1 Representing and Acquiring Thematic Roles

The concept of *thematic role* is one of the most vaguely defined, yet appealing, technical tools in the linguist's toolkit. Since its settlement in the circle of relevant modern theoretical issues, thanks to investigations by Tesnière (1959), Gruber (1965), Fillmore (1968), and Jackendoff (1972), this concept has been approached with what Dowty (1989) called the *I-can't-define-it-but-I-know-it-when-I-see-it* stance; that is, by using it without offering a proper definition. As easily predictable, such a state of affairs led to the proliferation of many alternative terms forged to refer to very close, if not identical, intuitions: *case relations*, *theta roles*, *semantic roles* and *thematic relations*. All these approaches share the general idea that thematic roles describe what can be intuitively depicted as the role played by an argument in the event or situation described by a verb, and little formalization has been obtained since early documented proposals such as Pānini's *kārakas*.

In natural language processing (NLP), thematic roles are both a valuable source of semantic knowledge encoded in lexical resources such as VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008), FrameNet (Baker et al., 1998), and PropBank (Kingsbury and Palmer, 2003), as well as the target of automatic extraction models, usually referred to as Semantic Role Labeling tools (see Gildea and Jurafsky, 2002; Lluís Márquez, 2008; Palmer et al., 2010). This information has proven useful for a variety of tasks, including machine translation (e.g., Liu and Gildea, 2010; Wu and Palmer, 2011) and question answering (e.g., Shen and Lapata, 2007). However, most of the computational linguistics literature still sees semantic roles as unanalyzable and unitary entities, thus relying on a view whose dramatic limitations have long been identified (for a review, see Levin and Rappaport Hovav, 2005).

A real breakthrough in the linguistic research on thematic roles was carried out by David Dowty. Rather than pursuing the impossible goal of defining an exhaustive taxonomy of semantic roles, Dowty argued that roles are not discrete and categorical entities, but have the same prototype structure of other types of concepts. Dowty (1989) proposed a *neo-Davidsonian* approach in which thematic roles are seen as a set of entailments of a predicate over its arguments, and thus characterized as second-order properties, i.e., as predicates of predicates. He also distinguished *individual roles* from *linguistic roles*. The former are verb-specific roles defined by the entailments associated with a particular verb argument: for instance, the *builder-role* is the set of all the properties and inferences we can conclude about *x* solely from knowing that *x builds y* is true. Linguistic roles are instead more abstract concepts shared among many verbs. Dowty (1991) assumed two basic linguistic roles, proto-agent and proto-patient, defined as a clusters of properties or entailments, organized like

the prototypes of Rosch and Mervis (1975). For instance, he described the proto-agent role as characterized by entailments such as *(volitional involvement in the event)*, *(sentience and/or perception)*, etc. Linguistic roles take on a special status in linguistic theory as they enter into grammatical generalizations, given that proto-agents and proto-patients tend to be realized as subjects and direct objects, respectively, in active sentences.

On the NLP side, work by Reisinger et al. (2015) shows that Dowty's proto-role hypothesis can be empirically validated by exploiting a large-scale crowd-sourced annotation task using corpus data. These authors then compared the results of the annotation they collected against those available in more conventional resources such as VerbNet. By building on encouraging results, finally, these scholars propose a novel task, Semantic Proto-Role Labeling, in which a system is asked to annotate a sentence with “scalar judgments of Dowty-inspired properties,” rather than with more conventional categorical thematic roles.

Acknowledging that thematic roles, both verb-specific and general, are to be conceived as clusters of properties entailed by verb arguments in turn raises two crucial issues that represent the main focus of this paper: (1) *How can we identify the specific entailments that characterize the thematic roles of a verb?* (2) *How do we learn the entailments associated with these thematic roles?* The first issue concerns the empirical evidence we can use to ground the study of thematic roles on a firm scientific foundation. McRae et al. (1997b) propose to identify the entailments of verb-specific roles using the features produced by a group of native speakers in a norming experiment. The feature-norming paradigm is in fact commonly adopted to investigate the content of conceptual knowledge in semantic memory. Because thematic role concepts are conceived as clusters of properties, subjects' elicited features can be used to identify information associated with the roles of specific events and to estimate its degree of prototypicality.

Concerning the way thematic roles are acquired, we endorse the claim by McRae et al. (1997b) that “role concepts are formed through the everyday experiences during which people learn about the entities and objects that tend to play certain roles in certain events” (p. 141). Similar to nominal concepts, thematic roles are organized in hierarchical structures leading from verb-specific roles to more abstract thematic concepts: for instance, the role of “accuser” is regarded as a subtype of the more general role of “agent.” Therefore, both individual and general roles result from an inductive process of abstraction from event knowledge. This is similar to the way roles are organized in FrameNet: verbs evoke event-specific frames that are part of an inheritance network whose top nodes correspond to abstract event schemas containing general roles like agent or patient. Psycholinguistic research has indeed provided robust evidence that

online sentence processing is deeply influenced by knowledge about events and their thematic roles (for a review, see McRae and Matsuki, 2009): verbs seem to be able to prime nouns describing the typical participant to the event they describe (Altmann and Kamide, 1999; Ferretti et al., 2001; Hare et al., 2009), especially in the presence of certain syntactic and grammatical cues (Traxler et al., 2001; Ferretti et al., 2001, 2007; Altmann and Kamide, 2007); nouns too appear to be able to prime both the other participants of an event (McRae et al., 1998; Kamide et al., 2003; Bicknell et al., 2010), as well as those verbs describing the events in which they typically participate (McRae et al., 2005a), a behavior that is useful to select a given verb sense (Matsuki et al., 2011). This knowledge is referred to by McRae and Matsuki (2009) as *Generalized Event Knowledge* because it consists of general encyclopedic information about the prototypical organization and unfolding of events. Generalized Event Knowledge is acquired through different sources, most importantly first-hand participation in events and language. For instance, the entailments characterizing the agent role of *accuse* derive from our experiences with people who accuse others and from linguistic descriptions of such events.

Our goal in this chapter is to investigate the contribution of language as a source of the entailments that characterize verb-specific semantic roles. In particular, we aim to explore to what extent the entailments activated by the thematic roles of a subset of English verbs can be acquired from their usage in a corpus. To address this question, we are proposing a distributional model in which the semantic content of the proto-agent and proto-patient role of a verb are characterized by the sets of verbs and nominal predicates that are strongly associated with them in texts. As an example, what our model looks after is the fact that the agent role of the target verb TO EAT can be described with properties such as *s/he drinks (while eating)*, that *s/he will digest what s/he has eaten*, and that *s/he was previously hungry*. In this preliminary work, this information will be represented by a strong association of this verb-specific role with verbs like *(to drink)* and *(to digest)*, as well as with adjectives like *(hungry)*.¹ We will test our model by comparing the extracted information against the properties produced by a group of native speakers to describe the content of verb-specific thematic roles.

The chapter is organized as follows. In Section 6.2 we illustrate a feature-norming study by means of which, following works by McRae et al. (1997b) and Lebani et al. (2015), we describe the thematic roles associated with twenty English verbs. In Sections 6.3 and 6.4 we show how a simple distributional model is apt to extract such information from a corpus, albeit with

¹ Throughout these pages, target verbs will be printed in SMALL CAPITAL font, whereas speaker-generated and automatically extracted descriptions will be enclosed in *(angle brackets)*.

a series of limitations and blindspots on which we will speculate in the final section.

6.2 Characterizing the Semantic Content of Verb Proto-roles

In the modern psycholinguistic literature, the feature norm paradigm has been widely employed to characterize the semantic content of the human conceptual knowledge. In its simpler form, it requires native speakers to produce short phrases to describe a set of target concepts. The collected descriptions are then normalized and categorized by the experimenter to build a dataset of pairings concept-feature of the form DOG *{has a tail}*, LOUNGE *{is fancy}*, or AIRPLANE *{flies}*.

Freely available resources built by exploiting different implementations of the feature norm paradigm are available for a limited number of languages, including English (Garrard et al., 2001; McRae et al., 2005b; Vinson and Vigliocco, 2008; Devereux et al., 2014), Italian (Kremer and Baroni, 2011; Lebani, 2012; Montefinese et al., 2013; Lenci et al., 2013), Dutch (De Deyne et al., 2008), and German (Kremer and Baroni, 2011; Roller and Schulte im Walde, 2014). These collections have been used as experimental stimuli (Ashcraft, 1978; Vigliocco et al., 2006), as a source of knowledge in proposing a model of semantic memory (Collins and Loftus, 1975; Hinton and Shallice, 1991; McRae et al., 1997a; Vigliocco et al., 2004; Storms et al., 2010), to investigate the pattern of impairments shown by anomic patients (Garrard et al., 2001; McRae and Cree, 2001; Vinson et al., 2003; Sartori and Lombardi, 2004), and in research on the nature of empirical phenomena such as semantic priming (Cree et al., 1999; Vigliocco et al., 2004), semantic compositionality (Hampton, 1979), and categorization (Smith et al., 1974; Rosch and Mervis, 1975).

In the computational linguistics literature, feature norms collections have been used to evaluate semantic extraction methods (Baroni et al., 2008; Baroni and Lenci, 2010) or as a source of semantic knowledge that can be exploited to enrich existing resources or other kinds of knowledge (Barbu and Poesio, 2008; Andrews et al., 2009; Steyvers et al., 2011; Lebani, 2012; Fagharasan et al., 2015). Some scholars even tested systems specifically tuned to extract feature-like semantic knowledge (Poesio et al., 2008; Devereux et al., 2009; Baroni et al., 2010; Kelly et al., 2010, 2013).

With few exceptions, most of the available feature norms have been collected for nominal concrete concepts expressed as nominal entities. One of these exceptions is the dataset assembled by Vinson and Vigliocco (2008), where 287 of the 456 described concepts denote actions, in 217 cases by means of a verbal lemma. In the paradigm adopted by these scholars there is no difference in the way verbs and nouns are collected and represented, so that the final dataset

represents a unitary space, whose suitability for modeling the human semantic memory has been proved by the same authors (Vigliocco et al., 2004).

Whereas Vinson and Vigliocco (2008) were interested in the properties of the event denoted by the verb, McRae et al. (1997b) collected the characteristics of the proto-agent and proto-patient roles for a group of 20 English transitive verbs, thus showing how the feature norm paradigm can be used to empirically characterize the semantic content of thematic roles. The scholars opted for a traditional paper-and-pencil setting, in which 32 participants were asked to list the characteristics of only one role for each verb. Crucially, instructions explicitly stated that what they were asked to list were not the typical fillers of a role (e.g., *judge* as the agent of TO CONVICT), but their characteristics (e.g., *(is old)* for the same role). McRae et al. (1997b) collected 1,573 distinct descriptions, 445 of which have been produced by 3 or more participants. Overall, no clear advantage of one proto-role over the other has been recorded, but the distribution of the features in the different verb-role pairs is far from uniform, a phenomenon that the authors ascribed to the well-known fact that some roles for some verbs admit a more restrictive group of fillers than others. Examples of highly consistent verb-specific roles include the proto-agent of the verb TO RESCUE and the proto-patient of the verb TO TEACH, whereas loosely defined roles include the patients of TO ACCUSE and TO SERVE.

By building on the observations by McRae and colleagues, Lebani et al. (2015) applied a modified version of this paradigm to a set of 20 Italian verbs. These authors modified the original methodology in several ways: by submitting the questionnaire online to a group of selected participants; by asking each participant to rate all possible verb-role pairs; by providing the participants with instructions to the form “*describe who CONVICTS*” or “*describe who IS CONVICTED*”, to avoid confronting the subjects with an elusive concept such as thematic role. The biggest modification to the original paradigm, however, was the explicit request to describe each role of each verb with respect to three different time slots:

- *before* the event described by the verb takes place: for instance, properties like *(to be ill)* for the patient of the verb TO CURE;
- *while* the event described by the verb takes place: for instance, properties like *(to speak)* for the agent of the verb TO TEACH;
- *after* the event described by the verbs took place: for instance, properties like *(to feel fine)* for the patient of the verb TO CURE.

Lebani et al. (2015) evaluated the impact of this last manipulation against an online reimplementation of McRae’s paradigm (McRae et al., 1997b), and the collected features set was much less skewed toward the required characteristics, and way more informative of the entailed properties that a filler acquires

by participating in the event described by a verb. It is this characteristic that drove our choice to adopt this last paradigm to collect, for 20 English transitive verbs, a description of the semantic content of the proto-agent and proto-patient semantic roles, to be later used as an evaluation benchmark against the neo-Davidsonian distributional model that we describe in the next section.

6.2.1 *Method*

To collect data from English native speakers, we crowdsourced our elicitation task through the Crowdflower marketplace.² Such a solution is usually adopted to collect a great amount of annotations or data, and to do so as quickly and cheaply as possible. But this often comes with the price of lower reliability and/or precision of the data due to the influence of many noncontrollable variables (on these topics, see Snow et al., 2008; Fort et al., 2011). Even if other authors proved that the collection of featural descriptions is a task that can be easily crowdsourced (e.g., Roller and Schulte im Walde, 2014), this required an adaptation of the procedure in Lebani et al. (2015) in order to submit our workers to a task that is not too labor-intensive and to filter unreliable data (see Kittur et al., 2013).

Materials We borrowed our experimental stimuli from McRae et al. (1997b). These were the following 20 English transitive verbs holding animate agents and patients: TO CONVICT, TO TEACH, TO RESCUE, TO ENTERTAIN, TO FIRE, TO CURE, TO PUNISH, TO HIRE, TO EVALUATE, TO ARREST, TO LECTURE, TO FRIGHTEN, TO INSTRUCT, TO TERRORISE, TO INVESTIGATE, TO WORSHIP, TO INTERVIEW, TO ACCUSE, TO SERVE, TO INTERROGATE.

The semantics of each possible verb-role-slot combination was then paraphrased to create requests of the form “please, list some of the features possessed by someone that [*inflected verb*] someone else” for the agent role and “please, list some of the features possessed by someone that is [*inflected verb*] by someone else.” For instance, the six requests created for the agent and patient role of the verb TO FIRE were, respectively: “please, list some of the features possessed by someone that [*fired* | *is firing* | *is going to fire*] someone else” and “please, list some of the features possessed by someone that [*has been fired* | *is being fired* | *is going to be fired*] by someone else”. Overall, 120 test questions of this sort were created, each to be used as the microtask to be submitted to our workers.

Procedure In each microtask the worker was requested to supply 5 to 10 short descriptions for a verb-role-slot triple. Microtasks were submitted

² Accessible at www.crowdflower.com

Describe The Features Of Someone Involved In An Event

Instructions ▾

Write 5 to 10 short sentences (one sentence per form) describing some features of a person involved in an event BEFORE, DURING or AFTER the event takes place.

example:

- person who HELPS someone else:
 - > [Before]: he is in danger; he cries for help; he is worried; he did something wrong
 - > [Before]: he is a kind person; he may be a stranger; he is on his own
 - > [During]: he may show off; he may be rude
 - > [After]: he may feel right; he may expect a reward; he should be rewarded

• person who IS HELPED:

- > [Before]: he is in danger; he cries for help; he is worried; he did something wrong
- > [Before]: he just watched; he feels relieved
- > [After]: he is grateful; he is safe; he feels better; he may feel guilty; he feels relieved

please, list some of the FEATURES possessed by someone that is GOING TO HIRE someone else

Feature 1

Feature 2

Feature 3

Feature 4

Feature 5

Feature 6

Feature 7

Feature 8

Feature 9

Feature 10

Can "enlist" and "hire" mean the same thing?

Yes

No

Can "hire" and "employ" mean the same thing?

Yes

No

Figure 6.1 The verb role description interface in Crowdflower.

by means of a web page similar to the one in Figure 6.1. The top of the page supplied intuitive instructions, along with exemplar descriptions for the verb TO HELP. The main area of the page, i.e., the one with a white background, presented the test question, followed by 10 empty forms and 2 language comprehension questions. Test questions required the worker to indicate whether the meaning of the target verb was similar to that of a test verb, which could be either a synonym of the target or a completely unrelated word. Each worker was free to complete from 1 to 120 different microtasks, presented in randomized order. On average, workers needed 116.02 s ($SD = 96.98$) to complete a valid hit. Hits were rejected if they met any of the following conditions:

- the worker didn't answer correctly to any test question;
- the worker completed the task in less than 30 s³;

³ Both the use of test questions and the duration threshold were intended to identify scammers. As a matter of fact, a manual inspection of the data showed that the latter strategy was more efficient than the former.

- the worker failed to provide at least three valid descriptions;
- the worker clearly misunderstood the requirements of the task.

The data collection process took place at the end of September 2014, and ended when 15 usable annotations for each verb-role-slot question was recorded, that is, after approximately 7 days.

Participants Eighty-seven unique workers contributed to the norming experiment, receiving €0.05 per hit. Only Crowdflower-certified “highest quality” contributors from the United Kingdom, the United States, or Ireland were allowed to participate. On average, each subject completed 20.7 ($SD = 26.64$) approved hits.

6.2.2 *Selection and Normalization*

The collected raw descriptions were first manually inspected to remove unwanted material such as incomplete sentences, meaningless descriptions, and all cases in which the worker reported the filler of the thematic role rather than its characteristics.

The selected descriptions were then “normalized,” that is, manipulated to identify meaningful chunks of information. Normalization practices in the literature can be organized into three main classes: minimal normalization (e.g., De Deyne et al., 2008); raw descriptions rewritten to conform to a phrase template (e.g., McRae et al., 1997b; Garrard et al., 2001; McRae et al., 2005b; Kremer and Baroni, 2011; Lenci et al., 2013; Devereux et al., 2014; Roller and Schulte im Walde, 2014; Lebani et al., 2015); or raw descriptions reduced to a list of focal concepts (e.g., Vinson and Vigliocco, 2008; Lebani, 2012). Common to virtually all strategies is a first step in which:

- spelling and orthography are standardized;
- conjoint and disjunct features are split: accordingly, a description such as *{is tasty and delicious}* should be split into *{is tasty}* and *{is delicious}*;
- auxiliaries and modal are stripped away: for instance, the description *{could be guilty}* should be simplified into a phrase like *{is guilty}*.

The main reason for us to collect featural descriptions was to evaluate a distributional model, so that the subsequent normalization steps were aimed at a – sometimes brutal – reduction of the raw description phrases into lists of focal concepts, a strategy analogous to those adopted by Vinson and Vigliocco

(2008) and Lebani (2012).⁴ In this second step, several crucial manipulations are performed:

- quantifiers are removed: for instance, the description *<has five legs>* can be simplified into something of the form *<has legs>*;
- the prominent concept(s) of each description are identified, and the remaining linguistic material discarded: for instance, two important chunks of information are available in the description *<has beautiful legs>*, thus leading to the creation of the two focal features *<beautiful>* and *<legs>*;
- the identified focal concepts are then lemmatized: e.g., plural nouns become singular, participles and gerunds are reported in their base form.
- synonymous features produced in different hits were encoded by using their most recurrent linguistic form: if two workers produced the description *<is calm>* and another produced *<is cool>*, then all descriptions were treated as instances of the same feature, i.e., *<is calm>*. Synonymous descriptions or focal concepts produced in the same hit, however, were analyzed as redundancies and discarded.

Slots merging and norms expansion The dataset of verb-role-slot features collected so far is analogous to the one described by Lebani et al. (2015), and throughout this chapter we refer to its features as *slot-based features*. For two reasons, however, this dataset is not optimal for our purposes, that is, to serve as a gold standard in the evaluation of our model. First of all, our model does not attempt to extract the temporal signature of each feature. The reason we resorted to this paradigm was its superiority in extracting “entailed” properties. We therefore merged all the feature sets produced for a given verb-role pair, irrespective of their temporal characterization. We refer to these as *role-based features*.

The second issue has been recognized and widely discussed in the relevant literature (e.g., Barbu and Poesio, 2008; Baroni et al., 2008; Baroni and Lenci, 2010). It pertains to the fact that the normalization process has the side effect of reducing the lexical richness of the uttered descriptions. When using a feature norm collection as a gold standard, lexical paucity has a direct impact on the evaluation statistics by artificially increasing the number of false negatives (i.e., properties extracted by the system but not linked to a synonymous description in the norms). In using the concrete concept properties of McRae and colleagues (McRae et al., 2005b) as a gold standard for the European

⁴ In fairness, different sets of norms have been prepared, each developed by following one of the three different normalization strategies. Given the scope of this chapter in these pages, we focus solely on those obtained by reducing the raw descriptions to their focal concepts.

Summer School in Logic, Language, and Information (ESSLLI) 2008 Distributional Semantic Workshop unconstrained property-generation task, Baroni et al. (2008) expanded their reference norms by (1) selecting the top ten features for each described concept, (2) extracting from WordNet (Fellbaum, 1998) the synonyms of each last word of each feature, and (3) performing a manual check to filter irrelevant synonyms and to add other potential linguistic material. Along these lines, we expanded our role-based features by extracting from WordNet (Fellbaum, 1998) all the synonyms of each of our focal concepts, without manual intervention. We refer to these as *expanded-role-based features*.

6.2.3 Results

Overall, our workers produced 11,985 raw descriptions, uniformly distributed along thematic roles (6,066 for the agent roles and 5,918 for the patient roles) and time slots (3,964 for the *before* slots, 4,016 for the *during* slots, and 4,004 for the *after* slots). Each hit returned, on average, 6.66 raw features ($SD = 2.17$). By splitting conjoint and disjunct descriptions the total climbs to 12,091, of which 392 were later discarded because they contained unwanted material or redundant information.

The normalization process resulted in 12,802 raw slot-based features. From these, 9,667 distinct verb-role-slot features were collected: 5,136 for the agent roles and 4,531 for the patient ones. In contrast with that reported by Lebani et al. (2015), this difference reaches statistical significance according to a paired Student's *t*-test ($t = 7.49$, $df = 19$, $p < 0.001$). On the other side, the distribution is pretty even across the different time slots: 3,190, 3,272, and 3,205 for the *before*, *during*, and *after* slots, respectively. A chi-square analysis failed to reveal any significant pattern both in the distribution of the features for slot both in the whole dataset ($\chi^2 = 0.57$, $df = 2$, $p > 0.1$), and among the two groups of thematic roles ($\chi^2 = 1.18$, $df = 2$, $p > 0.1$).

On average, each distinct slot-based feature has been produced by 1.32 ($SD = 0.945$) workers, and consistent features (those with frequency ≥ 2) accounts for the 17.69% of the total distinct slot-based features: 827 for the agent roles and 883 for the patient ones; 572, 563, and 575 for the *before*, *during*, and *after* slots, respectively. A paired Student's *t*-test shows that the difference in consistency between the two thematic roles reaches statistical significance, too ($t = -5.88$, $df = 19$, $p < 0.001$). A chi-square analysis failed to reveal any significant pattern both in the feature consistency of the time slots both in the whole dataset ($\chi^2 = 0.34$, $df = 2$, $p > 0.1$) and among the two groups of thematic roles ($\chi^2 = 0.14$, $df = 2$, $p > 0.1$).

Table 6.1 reports the number of featural descriptions and their consistency across the verb-role pairings. What it clearly shows is that, abstracting away

Table 6.1 *Distinct Slot-Based Features and Consistency for Verb-Role Pair*

	Agent		Patient	
	SB features	Consistency ^a	SB features	Consistency
TO ACCUSE	275	9.09%	214	19.63%
TO ARREST	259	18.15%	237	18.14%
TO CONVICT	276	16.3%	220	20.91%
TO CURE	235	20.0%	192	23.44%
TO ENTERTAIN	219	19.63%	209	23.44%
TO EVALUATE	275	13.09%	253	15.42%
TO FIRE	260	18.08%	256	17.97%
TO FRIGHTEN	241	16.6%	207	19.32%
TO HIRE	244	17.21%	198	23.74%
TO INSTRUCT	245	16.33%	248	17.34%
TO INTERROGATE	261	13.41%	236	17.8%
TO INTERVIEW	270	18.15%	231	20.78%
TO INVESTIGATE	271	16.97%	239	16.74%
TO LECTURE	287	14.63%	243	17.28%
TO PUNISH	243	18.93%	227	23.79%
TO RESCUE	218	20.64%	206	22.33%
TO SERVE	271	17.34%	214	20.09%
TO TEACH	243	17.7%	208	21.63%
TO TERRORIZE	279	8.96%	244	15.98%
TO WORSHIP	264	14.02%	249	17.67%

^a Consistency = the percentage of distinct features produced by two or more workers.

from the main opposition between agent and patient role, some thematic roles for some verbs are clearly better defined than others. For instance, just compare the consistency rate of the agent roles of TO ACCUSE and TO TERRORIZE with those for the verbs TO CURE and TO RESCUE. McRae et al. (1997b) see lack of consistency as a by-product of the fact that those roles can be realized in many possible ways: i.e., the range of people who typically *accuse* or *terrorize* is more varied than those who *cure* or *rescue*.

Gold standards Table 6.2 summarizes the distribution of distinct features for each verb-role pairing in the role-based and role-base-expanded datasets, i.e., in the two collections that will be used as gold standard for the evaluation of our model.

By removing the temporal characterization from our slot-based norms, i.e., by aggregating the features produced for each verb-role pairing, we obtained a total of 7,290 distinct role-based features. Of these, 3,923 were associated with an agent role and 3,367 with a patient role. The difference between the average

Table 6.2 *Distinct features in the gold standard datasets*

	Agent Features		Patient Features	
	Role-based	RB expanded	Role-based	RB expanded
TO ACCUSE	213	1,759	166	1,546
TO ARREST	190	1,481	180	1,677
TO CONVICT	204	1,693	157	1,138
TO CURE	175	1,184	142	1,204
TO ENTERTAIN	176	1,320	156	1,223
TO EVALUATE	226	1,788	187	1,650
TO FIRE	198	1,633	188	1,611
TO FRIGHTEN	193	1,629	150	1,205
TO HIRE	180	1,534	141	1,222
TO INSTRUCT	185	1,488	197	1,710
TO INTERROGATE	196	1,615	182	1,498
TO INTERVIEW	198	1,689	174	1,365
TO INVESTIGATE	208	1,643	173	1,255
TO LECTURE	227	1,912	181	1,633
TO PUNISH	182	1,619	166	1,429
TO RESCUE	155	1,488	150	1,462
TO SERVE	203	1,717	161	1,298
TO TEACH	191	1,586	146	1,116
TO TERRORIZE	224	1,696	182	1,309
TO WORSHIP	199	1,480	188	1,260

number of agent features produced for each verb ($M = 196.15$, $SD = 18.31$) and the average number of patient features ($M = 168.35$, $SD = 17.21$) reaches statistical significance ($t = 7.15$, $df = 19$, $p < 0.001$).

By automatically expanding our features with synonyms available in WordNet, we put together a dataset composed by 59,765 distinct expanded-role-based features, 31,954 for the agent roles and 27,811 for the patient ones. On average, each verb is associated with 1,597.7 agent ($SD = 164.05$) and 1,390.55 patient features ($SD = 194.29$). A paired Student's t -test reveals that such difference is significant ($t = 4.33$, $df = 19$, $p < 0.001$).

6.3 A Distributional Model of Thematic Roles

In computational linguistics, the concept of thematic role is often evoked when referring to two intercorrelated branches of research: the design of lexical resources (e.g., VerbNet, FrameNet, and PropBank), each typically implementing a different idea of what a role is, and the development of tools apt to annotate a sentence with the roles fulfilled by the verbs arguments, given a predefined list of semantic frames or thematic role labels.

Differently from these mainstream approaches, and in line with the work by Reisinger et al. (2015), we adopt a neo-Davidsonian perspective (i.e., we view roles as second-order properties), and we do not see thematic concepts as primitive entities, but as verb-specific concepts represented as clusters of features organized in a prototypical fashion (Dowty, 1991; McRae et al., 1997b). Our assumption in this chapter is that the features entering into the definition of thematic roles depend on the generalized knowledge about the events expressed by verbs. In particular, we argue that important aspects of such knowledge depend on the way verbs are used in linguistic contexts, and that therefore they can be modeled with distributional information automatically extracted from corpora. We are thus dealing with a problem of automatic lexical acquisition, which we tackle in an unsupervised manner, by relying on the minimal possible number of assumptions. Our aim is to present a computational model to extract from corpora the features characterizing verb-specific roles, which we test on the norms presented in Section 6.2. In this section, we review useful insights we borrowed from related literature on distributional semantics (Section 6.3.1) and on the automatic extraction of event chains from corpora (Section 6.3.2), and we present a short description of the core aspects of our model (Section 6.3.3).

6.3.1 Thematic Information in Distributional Semantics

Unsupervised corpus-based models of semantic representation (Sahlgren, 2006; Lenci, 2008; Turney and Pantel, 2010), commonly labeled as vector/semantic/word spaces or distributional semantic models (DSMs), have been established in the last thirty years as a valid alternative to traditional supervised and semisupervised methods. Among the many factors contributing to this success, probably the most cited is the fact that these models are faster and less labor-demanding than manual annotation and semisupervised models.

Another key factor, crucial for the work we present here, is that such models do not need prior knowledge other than that required to implement the so-called Distributional Hypothesis (Harris, 1954; Miller and Charles, 1991). This hypothesis has been received in the NLP literature as a working assumption roughly stating that the similarity of the contexts in which two linguistic expressions occur is a measure of their similarity in meaning (see Sahlgren, 2008, for a more in-depth discussion). This, in turn, is the corollary of another working assumption: that the meaning of a linguistic item is reflected in the way it is used.

Implementation of the distributional hypothesis depends on a few vaguely defined concepts, and the whole literature on DSMs is centered on the characterization of these concepts:

- *Linguistic expressions*: What kind of linguistic expressions can be characterized in distributional terms?
- *Context*: What is the most effective way to characterize the linguistic behavior of our target expressions?
- *Similarity*: How can we compare the linguistics contexts and what kind of semantic similarity can we model?

All existing DSM models incarnate alternative answers to such issues. Restricting this quick summary to DSMs representing words (see Turney and Pantel, 2010, for an overview of the possible target expressions), typically these models are built by scanning a corpus for all occurrences of the target expressions, identifying their contexts, and representing the words by context frequencies in a co-occurrence matrix. Contexts can be windows of words, syntactic relations, patterns of parts of speech, chapters, documents, and so forth (see Sahlgren, 2006, for a comparative review). Generally, the raw co-occurrence matrix is manipulated by (1) weighting the frequencies for highlighting meaningful word-context associations and (2) reducing dimensionality to create dense vectors of latent features for ignoring unwanted variance and/or for computational efficiency reasons. Each vector in the final matrix is assumed to represent the distributional signature of a target word, and is used to calculate the similarity with all the other words of interest according of a chosen vector similarity measure, typically the cosine. (For a critical overview of the commonly adopted technical solutions, see Bullinaria and Levy, 2007, 2012; Lapesa and Evert, 2014.)

Even if the DSM we present in these pages mostly conforms to this general pattern, to the best of our knowledge, no previous system has been proposed to extract the kind of information we are interested in. DSMs have been widely used for the SRL task (e.g., Erk, 2007; Collobert et al., 2011; Zapirain et al., 2013; Hermann et al., 2014; Roth and Lapata, 2015), but mostly to enhance the performance of a SRL classifier, as an ancillary source of information for a task based on a concept of semantic role that is incompatible with the one adopted in these pages.

Our model is directly inspired by works exploiting a distributional space in which linguistic expressions are characterized on the basis of the syntactic environment in which they occur, that is, syntax-based DSMs (e.g., Grefenstette, 1994; Lin, 1998; Padó and Lapata, 2007; Baroni and Lenci, 2010). In these models, syntactic environments are obtained by extracting from shallow-processed or full-parsed text dependency paths such as those linking a verb to its subject or its object. For instance, given the sentence *the supermodel left the catwalk*, in a syntax-based model the distributional entry for the verb TO LEAVE is enriched with a reprocessing of the dependency:filler patterns `subj:supermodel` and `obj:catwalk`. Syntax-based DSMs have proved

to be useful in many semantic tasks. However, the branch of research that uses such DSMs to model thematic fit is the most similar to ours, for two reasons: First, our work and that reviewed in the next subsection share the same view on the usage-based nature of thematic roles; moreover, we all adopt the working assumption that syntactic slots can be seen as rough approximation of semantic roles, at least in a corpus-based model.

The concept of *thematic fit* refers to the appropriateness of a lemma as a filler of a given verb-specific thematic role for a verb. The cognitive relevance of this notion has been widely proved and tested in psycholinguistics (for a review, see McRae and Matsuki, 2009), where thematic fit judgments are typically collected by asking speakers to rate the plausibility of a lemma being a filler of a given thematic role for a given verb. Such a notion is intimately related to, although not equivalent to, the notion of selectional preference, the main difference being the nature of the involved elements: discrete semantic types in the case of selectional preferences, gradient compatibility of an argument with a thematic role in the case of thematic fit.

To the best of our knowledge, Erk et al. (2010) were the first to evaluate a syntax-based DSM against human-generated thematic fit judgments. In the exemplar model described by these scholars, i.e., the EPP model first introduced by Erk (2007), plausibility scores for argument filler are computed by measuring the similarity of the new candidates with all the previously attested fillers for that verb-role pairing. Crucially, the distributional knowledge extracted in this model comes from two corpora, or from different uses of the same corpus: a “primary” corpus, used to obtain information about verb-argument co-occurrences, and a “generalization” corpus, exploited to extract similarity measures between argument fillers. Erk and colleagues tested their proposal by correlating the plausibility values produced by the system against the human-generated judgments collected by McRae et al. (1998) and those by Padó (2007). The crucially different sparsity degrees of the stimuli in the two datasets clearly affected the performance of the model, which, all things considered, mildly correlated with human judgments only on the latter dataset.

Similar results, with slightly higher correlations, were reported by Baroni and Lenci (2010) when evaluating their framework, Distributional Memory (DM), against the same judgments. In contrast to the practice of developing different DSMs for different tasks, DM is a framework in which co-occurrence information is extracted just once and represented into a third-order tensor that functions as a semantic knowledge repository. When tackling a specific task, the DM tensor is then manipulated to create the task-specific DSM as needed, without resorting back to the corpus. In modeling thematic fit, Baroni and Lenci (2010) showed how their tensor can be manipulated to derive a matrix in which the vectors are the target lemmas and the dimensions are dependency : filler patterns. This syntactic DSM, analogous to the representation exploited by Erk

et al. (2010), is then used to identify, for each verb, its typical subject and object fillers, to built their centroids (i.e., their “prototypical” vectors), and to predict thematic fit for a given noun-role-verb by measuring the distance between the target noun and the verb-role centroid.

Greenberg et al. (2015) compared the performance of the model by Baroni and Lenci (2010) with those that can be obtained by two different role-based DSMs. Moreover, they experimented with different methods to calculate the prototypical vector set for each verb-role. The results of this comparative work, evaluated against the datasets by McRae et al. (1998) and by Padó (2007), together with the instrument and location roles judgments by Ferretti et al. (2001), showed a slight superior performance for the DM-based model,⁵ and a clear constant improvement in using agglomerative clustering to build the prototypical filler of a verb-specific role. Finally, Lenci (2011) goes further in the investigation of the thematic fit phenomenon by using the same DM-derived matrix as Baroni and Lenci (2010) to model how argument expectations are updated on the basis of the realization of the other roles in the verbs argument structure. Evaluated against data from Bicknell et al. (2010), the best settings of this model obtained a 73-84% hit accuracy rate.

Another strand of research that has been inspirational for our work includes those works that try to model feature norms information for concrete concepts by means of a DSM. The first attempts to automatically extract short descriptions of this sort are described in Almuhabeb and Poesio (2004, 2005) and Barbu (2008). These approaches were quite limited in their scope, being focused on a restricted set of semantic relations, two in the former studies, six in the latter. To the best of our knowledge, Baroni et al. (2010) were the first to tackle an unconditional version of this task. Their model Strudel extracts properties by looking at the distribution of superficial patterns like [Concept]_is_ADV_[Property] (as in *the grass is really green*) or [Property]_of_[Concept] (as in *pack of wolves*). The key intuition is that a strong semantic link between a concept and a property reflects in their co-occurrence in a great variety of different patterns. Evaluated against the ESSLLI dataset, the authors reported a precision score of 23.9%, to date the highest score registered for the unconstrained extraction of feature-like $\langle \text{concept}, \text{property} \rangle$ pairs. As argued by Devereux et al. (2009), a major limitation of the Strudel approach is that the semantic relations between concepts and properties are characterized only implicitly, i.e., by means of superficial patterns. In fact, Strudel can

⁵ For the sake of completeness, it should be noted that Erk et al. (2010) also compared the results obtained by exploiting, as a primary corpus, a role-semantic rather than a syntactic annotation, and report a slight advantage of the former over the latter. As noted by the authors, however, the presence of the many sources of variance (manual vs. automatic annotation, corpus size) doesn't allow any firm conclusion from these results.

be seen as an unconstrained model to extract feature-like $\langle \text{concept}, \text{property} \rangle$ pairs.

Devereux, Kelly, and colleagues (Devereux et al., 2009; Kelly et al., 2010) were the first scholars to try to automatically extract feature-like $\langle \text{concept}, \text{relation}, \text{property} \rangle$ triples. They tried to identify the prototypical properties of a concept and to explicitly characterize the type of their relation. The model they proposed articulates in two phases: first, manually generated syntax-based rules were used to extract a set of candidate $\langle \text{concept}, \text{relation}, \text{property} \rangle$ triples; then these triples were ranked on the basis of the conditional probabilities of concept and feature classes derived from the McRae dataset. As reported by Kelly et al. (2010), when evaluated against the ESSLLI dataset, their best model obtained a precision score of 19.43% for the identification of $\langle \text{concept}, \text{property} \rangle$ pairs and 11.02% when looking for $\langle \text{concept}, \text{relation}, \text{property} \rangle$ triples.

Kelly et al. (2013) moves on by proposing a model that exploits syntactic, semantic, and encyclopedic information. This model starts by applying a series of rules to extract meaningful paths from the syntactic annotation available in two corpora: an encyclopedic corpus and a general corpus. Then the model weights each candidate triple first by using a linear combination of four metrics and later applying the same reweighting strategy as in Devereux et al., 2009; Kelly et al., 2010. When evaluated against the same settings used by Baroni et al. (2010), their best models obtain a precision score of 13.39% for the identification of $\langle \text{concept}, \text{property} \rangle$ pairs and 5.02% when looking for $\langle \text{concept}, \text{relation}, \text{property} \rangle$ triples.

6.3.2 A Wider Context: Narrative Event Chains

Another branch of research investigating an issue related to ours focuses on the unsupervised characterization of *Narrative Event Chains*, defined as partially ordered set of events involving the same protagonist (Chambers and Jurafsky, 2008), where an event is represented by a verb together with its arguments. The following example, adapted from Chambers and Jurafsky (2009), describes a chain in which the protagonist is being prosecuted. The sequence of the events in this chains can be summarized as: the protagonist admits something and pleads (guilty), before being convicted and sentenced. Formally, this chain can be represented as a tuple (L, O) , where L is a set of $\langle \text{event}, \text{argument slot} \rangle$ tuples and O is a partial temporal ordering:

$$L = \langle \text{admit, subject} \rangle, \langle \text{plead, subject} \rangle, \langle \text{convict, object} \rangle, \langle \text{sentence, object} \rangle$$

$$O = \{(\text{plead, convict}), (\text{convict, sentence}), \dots\}$$

The unsupervised characterizations of event chains and related issues, such as the induction of event schemas and the temporal ordering of events, have

been tackled by relying on different source data, e.g., text corpora (e.g., Chambers and Jurafsky, 2008, 2009; Chambers, 2013; Balasubramanian et al., 2013) vs. crowdsourced descriptions (e.g., Regneri et al., 2010; Frermann et al., 2014), and on different families of approaches, e.g., graph-based methods (e.g., Regneri et al., 2010; Balasubramanian et al., 2013), probabilistic approaches (e.g., Cheung et al., 2013; Chambers, 2013; Frermann et al., 2014) or distributional learning (e.g., Chambers and Jurafsky, 2008, 2009).

There is a close relationship between the concept of event chain and the entailment-based concept of semantic role we adopt in this chapter. In a sense, part of the verb-specific entailments we aim to model is what happens to a protagonist (i.e., the role filler) in a prototypical event chain if we take our target verb as a reference point. As an example, let us go back to the prosecution narrative chains mentioned earlier and suppose that we have proved that they describe a prototypical sequence of events. At least some of the entailments associated to the patient of the verb TO CONVICT correspond to what happens to her/him before, during, and after the conviction event takes place. These entailed actions and properties may be found among the events that compose a prototypical narrative schema containing our target *(event, argument)* pairing: for instance, *s/he admits*, *s/he pleads (guilty)*, and *s/he is convicted*.

With this parallelism in mind, we looked at the seminal models by Chambers and Jurafsky (2008, 2009) for useful insights and intuition to integrate into our model, especially in light of the methodological affinities between our works. The starting point of Chambers and Jurafsky (2008) is the “narrative coherence” assumption: verbs whose arguments belong to the same coreference chain are semantically connected, and more likely to participate in a narrative chain. Briefly, the model proposed by these authors articulates in three steps. In the first step, the protagonist and the subevents are identified by first calculating the strength of association between pairs events, where the association score is a function of how often two events have a coreferring entity, and combining these pairwise associations into a global narrative score. Evaluated with a variation of the cloze task (Taylor, 1953), such a method shows a 36% improvement over baseline. Association scores are later fed to an agglomerative clustering algorithm to construct discrete narrative chains. In parallel, a two-stage machine learning architecture is used to temporally order these connected subevents, obtaining a 25% increase over a baseline for temporal coherence.

Chambers and Jurafsky (2009) extended these results by dealing with two limitations of their previous proposal: the lack of information concerning the role or type of the protagonist and the fact that only one participant was represented. As a solution to the former issue, the authors propose the notion of “typed” narrative chains, that is, an extension of the notion of chain in which the argument shared between events is defined by being a member of a given set

of lexical units, nouns cluster, or other semantically motivated group. The second extension results in the introduction of the concept of “narrative schema,” that is, an extension of the notion of narrative chain that models the entire narrative of the document by generalizing over all the actors involved in a set of events. When tested against the same dataset of Chambers and Jurafsky (2008), the joint effect of both extensions resulted in a 10% increment over the performance of the previous model.

6.3.3 *The Core of a Neo-Davidsonian DSM for Semantic Roles*

In the rest of this section we describe the core characteristics of a DSM incorporating a neo-Davidsonian view of verb-specific roles as clusters of prototypical features derived from corpus-based distributional data. In the next section we describe how we translated this model into an algorithm that we tested against the human-elicited properties described in Section 6.2.

Our main assumption is that (at least a subset of) the entailments associated with the specific roles of a target verb derive from the actions and properties associated with the role fillers in prototypical narrative schemata containing our target verb. Given a verb v and its specific role r_v , we define f_1, \dots, f_n as the n -most prototypical noun fillers of r_v : for instance, if r_v is the patient role of TO CONVICT, the fillers can be *defendant*, *prisoner*, etc. Let s_1, \dots, s_n be the narrative sequences of events in which the role-filler pairs $\langle r_v, f_i \rangle$ occur in a corpus. Each sequence s_i can be regarded as a broader scenario including the event expressed by the target verb v and the filler f_i for the role r_v . We then provide the following distributional characterization of verb-specific thematic roles:

The verb specific role r_v is the set of the predicates most associated with its fillers f_1, \dots, f_n in the narrative sequences s_1, \dots, s_n .

This framework thus relies on insights derived from both strands of research outlined in this section: from Erk et al. (2010) and subsequent works we borrowed the idea that thematic fit can be modeled by means of a syntax-based DSM; from Chambers and Jurafsky (2008) and subsequent works we borrowed the idea that the discourse structure, and in particular coreference chains, can be used to model sequences of events belonging to larger scripts or scenarios. In the final model, the semantic content of each verb-specific thematic role is represented by the set of predicates that meet the following two conditions:

- they are strongly associated with the prototypical fillers f_1, \dots, f_n of a verb-specific role r_v ;
- one of its argument frequently belongs to the same coreference chain as the filler of r_v .

These sets of entailments are identified by combining two contextual representations: a distributional syntax-based representation and a coreference-based representation. The distributional contextual representation is built in a three-step process:

1. A dependency extraction phase, during which a corpus is scanned to identify and manipulate all the relevant syntactic relations headed by a verb v . As noted by other authors (e.g., Preiss et al., 2007), such a process should be carefully tuned on the behavior of the specific parser used to annotate the input corpus.
2. Syntax-based co-occurrence frequencies are then used to calculate the association score between each verb-slot pairing and its fillers f_1, \dots, f_n . Given the symmetrical nature of association measures, this information can be used to model both direct and inverse selectional preferences (Erk et al., 2010). Accordingly, this step is used to select, for each verb-specific role r_v , its prototypical fillers as well as, for each filler, the prototypical verb-specific role in which it occurs. In what follows, we use the notation `relation-1: predicate` to refer to a construction representing both the inverse relation linking a lemma to its head, as well as the head. Its intuitive meaning can be paraphrased as “(the filler) is the *relation of predicate*,” e.g., `obj-1:eat` indicates a filler (e.g., *apple*) is the object argument of the verb `TO EAT`.
3. Finally, direct and inverse preferences are manipulated to associate each target r_v with a set of contextual `relation: predicate` constructions, obtained by interpreting the inverse selectional preferences of each r_v prototypical filler as clues of semantic relatedness. As an example, let us suppose that the target r_v is the patient role of the verb `TO WRITE` and that in the previous step we learned that its top associated fillers are *letter* and *book*. Let us assume that these nouns are strongly associated with the object position of {`TO RECEIVE`, `TO SEND`, `TO COMPOSE`} and {`TO READ`, `TO PUBLISH`, `TO DEDICATE`}. In this step we would elaborate on this picture to identify a set of candidate entailments such as the one between the patient of `TO WRITE` and `obj-1:publish` (i.e., what is written is typically published), or `TO WRITE` and `obj-1:read` (i.e., what is written is typically read).⁶

The distributional contextual representation collects events and properties that are related to our target verbs, but only a part of these are entailment patterns that may reasonably be assumed to enter into the definition of verb-specific thematic roles. For instance, while it is fairly plausible to presume the

⁶ Note that we are focusing solely on the object position of a verb just for the sake of exposition. As will become clear in the following section, this line of reasoning applies to all semantic roles we may find useful.

existence of an entailment relation between TO WRITE and TO PUBLISH, the relation between TO WRITE and TO COMPOSE is clearly one of near-synonymy. In fact, a crucial assumption of our model is that the distributional features characterizing verb roles belong to the event sequences including the target verb and its fillers. This is indeed the case of TO WRITE and TO PUBLISH, which can be assumed to be part of a larger book production scenario. We identify sequences of events including both the target verb and the extracted distributional context information with the following procedure of coreference-based contextual representation:

1. We extract from a coreference-annotated and parsed corpus all the verbs and nominal predicates whose argument typically belongs to the same coreference chains of the fillers f_1, \dots, f_n of our target verbs. Crucially, in this passage we keep track of the syntactic relation between each verb and the entities involved in each coreference chain. For instance, the text *I wrote you a note the other day. Did you read it? Yes, and I posted it online* contains chains linking the object of our target verb TO WRITE, the object of the verb TO READ, and the object of the verb TO POST.
2. From each coreference chain, we extract, for each target verb-specific role r_v (e.g., the object role of the verb TO WRITE), all the inverse dependencies involving each of the entities that corefer with our target verbs filler. In our example, this means isolating the contextual constructions obj-1:read and obj-1:post , jointly meaning something like “(the filler of our target verb-specific role corefers with) the object of the verb TO READ and the object of the verb TO POST.”
3. For each r_v , we removed the inverse dependencies missing from the distributional contextual representation and use the filtered coreference-based co-occurrence frequencies to calculate the strength of association between each target r_v (e.g., the object role of the verb TO WRITE) and each contextual construction (obj-1:read and obj-1:post). The most associated constructions are precisely the distributional features we use to characterize the entailments associated with r_v .

6.4 Experiments with Our Neo-Davidsonian Model

We tested the validity of our approach by evaluating how many of the speaker-generated entailment patterns collected in the experiment described in Section 6.2 we are able to automatically extract from an annotated corpus. To test the relative strength of the different sources of information, three different DSMs were created: the full model, implementing all the passages described in Section 6.3.3; a coreference model, in which only coreference-based

information were used; a distributional model, based solely on distributional contextual representations. We dubbed the latter two models *quasi-Davidsonian*.

All models were trained on a coreference-annotated and parsed version of the British National Corpus⁷ (BNC; Aston and Burnard, 1998), a 100M words corpus of British English language productions from a wide range of written (90%) and spoken (10%) sources, built in the first half of the 1990s. The corpus has previously been POS-tagged and lemmatized with the TreeTagger⁸ (Schmid, 1994), parsed with MaltParser⁹ (Nivre et al., 2007), and coreference-annotated with BART¹⁰ (Versley et al., 2008).

In collecting our gold standard role descriptions, we followed the settings of McRae et al. (1997b) which focused solely on the agent and patient proto-roles. Consequently, the following experiments address only these two roles. To limit data sparsity, we included in our test set only those verbs of McRae's list that occurred in the BNC more than 1,000 times, a condition that was not met by three verbs: TO TERRORISE ($f = 115$), TO INTERROGATE ($f = 274$), and TO WORSHIP ($f = 369$).

6.4.1 The Full Neo-Davidsonian Model

Implementation of the full model follows the general picture outlined in Section 6.3.3. As a first step, we scanned the syntactic annotation of the corpus and applied a set of parser-specific rules to handle such problematic phenomena as conversion of the passive diathesis, identification of the antecedents of relative pronouns, treatment of conjunct and disjunct; and identification and treatment of complex quantifiers such as “a lot of.”

We then looked in the corpus for instances of relevant syntactic relations, and for each occurrence we identified the element heading the relation and the lemma filling the argument position, thus obtaining a tuple of the form: $\langle \text{verb}, \text{relation}, \text{filler} \rangle$. In these experiments, the set of dependency relations we are interested in is composed by:

- sbj: *the professor wrote the letter* → $\langle \text{write}, \text{sbj}, \text{professor} \rangle$;
- obj: *the professor wrote the letter* → $\langle \text{write}, \text{obj}, \text{letter} \rangle$;
- prd: *the letter became famous* → $\langle \text{letter}, \text{pred}, \text{famous} \rangle$.

Following Erk et al. (2010), we see the sbj and obj relations as surface approximations of the agent and patient proto-roles. As such, they will be used both for characterizing the selectional preferences of our target verbs, as well as the inverse selectional preferences of their prototypical fillers. The prd relation,

⁷ www.natcorp.ox.ac.uk ⁸ www.cis.unit-muenchen.de/~schmid/tools/TreeTagger/

⁹ www.maltparser.org ¹⁰ www.bart-coref.org

on the other hand is only used to extract the inverse selectional preferences of the prototypical fillers of our target verbs. This way, we extract all those properties that are typically described by adjectives or nouns.

We used the frequency of each $\langle \text{verb}, \text{relation}, \text{filler} \rangle$ tuple to calculate the strength of association between verb-specific roles (i.e., $\langle \text{verb}, \text{relation} \rangle$ pairs) and fillers. In our experiments we used positive Local Mutual Information (pLMI) to calculate the strength of association between a target entity (e.g. a verb-specific role r_v) and a given context (e.g. a contextual construction). Local Mutual Information (LMI; Evert, 2009) is defined as the log ratio between the join probability of a target t_i and a context c_j and their marginal probabilities, multiplied by their joint frequency:

$$\text{LMI}(t_i, c_j) = f(t_i, c_j) * \log_2 \frac{p(t_i, c_j)}{p(t_i) * p(c_j)} \quad (6.1)$$

LMI is a version of the Pointwise Mutual Information (PMI; Church and Hanks, 1991) between a target and a context weighted by their joint frequency, usually preferred to PMI to avoid its characteristic bias toward low-frequency events. Positive LMI is obtained by replacing all negative values with 0:

$$p\text{LMI}(t_i, c_j) = \max(0, \text{LMI}(t_i, c_j)) \quad (6.2)$$

We used these statistics to select:

- *Direct selectional preferences*: The top 50 fillers associated with each target $\langle \text{verb}, \text{relation} \rangle$ tuple, where the relation can be either sbj or obj. For each target verb, therefore, we collected 50 subject and 50 object fillers;
- *Inverse selectional preferences*: The top 100 $\langle \text{verb}, \text{relation} \rangle$ tuples associated with each filler, where the relation can be either sbj, obj, or prd.

The distributional-based contextual representation is built by associating each target verb-specific roles r_v with the top $\langle \text{verb}, \text{relation} \rangle$ tuples of their top fillers. Alternatively said, the output of this phase will be obtained by merging, for each r_v , the inverse preferences of its prototypical fillers, thus obtaining a set of `relation-1:verb` contextual constructions.

In a second phase, for each r_v , we parsed all the coreference chains involving its fillers and isolated the verbal head or predicate of each entity in a ± 2 -sentences window that belongs to the chain of our verbs filler. We chose to focus on a portion of the coreference chain centered on the target verb to avoid those events of the narrative chains that are not directly related to our target event. We leave to future investigation the evaluation of the effect of this hyperparameter. In this passage, we track the dependency relations between the coreferring entities and their heads, thus obtaining sets of `relation-1:verb` contextual constructions analogous to the ones exploited as contexts in the distributional representation.

The two sets of contexts, i.e., the one used in the distributional model and the one used in the coreference model, are indeed comparable notwithstanding their different natures. They both encode different kinds of semantic relatedness: relatedness due to the sharing of the same sets of fillers in the case of the distributional contexts; relatedness due to the participation to the same event chains in the case of the coreference contexts. We take advantage of this compatibility by filtering the latter on the basis of the former. That is, for each verb-specific role r_v , we retain only the `relation-1:verb` contextual constructions that are shared between the distributional and the coreference representations. Finally, we resort to the coreference chains to extract the co-occurrence frequency between the r_v and the selected contextual constructions, and to calculate their association with pLMI. In our view, these top associated contextual constructions provide a distributional representation of the entailment patterns licensed by our verb-specific roles.

Table 6.3 in Appendix 1 reports the top associated contextual constructions that our model extracted for the agent and patient roles of the verbs TO ARREST and TO PUNISH. Intuitively, the high association between the agent role of the verb TO ARREST and the contextual constructions `sbj-1:hold` and `sbj-1:imprison` could be paraphrased as *s/he who arrests someone also holds him/her/someone else* and *s/he who arrests someone also imprisons him/her/someone else*. On the other hand, the high association between the patient role of the verb TO PUNISH and the contextual constructions `obj-1:torture` and `sbj-1:desperate` could be paraphrased as *s/he who is punished may be also tortured* and *s/he who is punished may be desperate*.

6.4.2 Quasi-Davidsonian Models

To evaluate the relative importance of the two souls of our neo-Davidsonian DSM, i.e., the distributional and the coreference-based components, we created two different DSMs, each modeling exclusively one kind of information. In the distributional model, the association between r_v and the contextual constructions is calculated solely on distributional basis, without trying to account for the patterns in the narrative structures that can be extracted from the coreference annotation. A coreference-based model relies solely on the information that can be extracted from the coreference chains, without resorting to syntax-based distributions to filter out infrequent fillers.

6.4.3 Evaluation

DSMs are usually evaluated extensionally, that is, by recording their performance on tasks that are supposed to tackle some crucial aspects of the human semantic memory. Typical tasks are to mimic the intended behavior of the

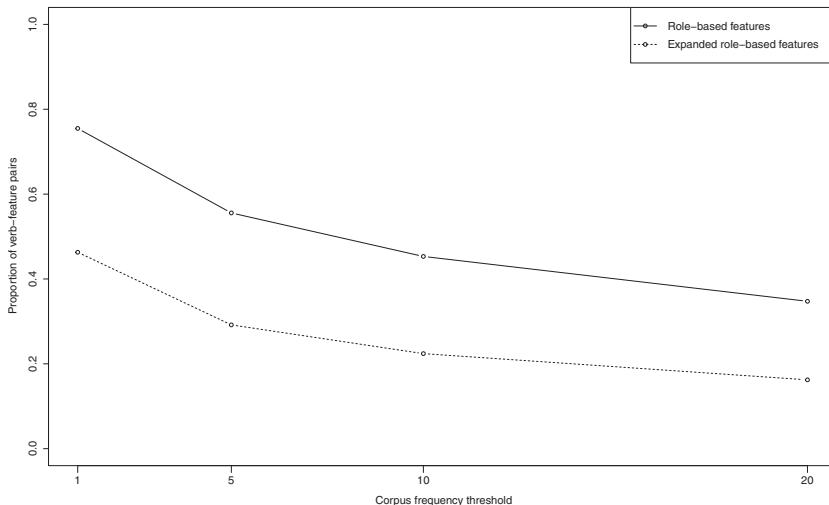


Figure 6.2 Proportion of verb-feature pairs in a ± 2 -sentences window, modulated by application of different frequency thresholds.

proficient speaker in tests like the synonym test questions from the Test of English as a Foreign Language the correlation with human-generated linguistic metajudgments, and clustering words to emulate some available semantic classification or how well they serve as features for machine learning algorithms.

Intensional methods, in which the validity of the semantic knowledge encoded in a DSM is directly assessed, are less common. Practices of this sort include the manual (usually crowdsourced) evaluation of the nearest neighbors returned by DSMs for target items, or the test against a prepared dataset of valid target-context associations such as speaker-elicited features.

In these pages we adhere to this latter tradition and evaluated our model against the two gold standard datasets described in Section 6.2.3, i.e., the role-based dataset obtained by stripping the temporal characterization from the descriptions collected in Section 6.2, and the expanded-role-based datasets built by enriching the role-based features with synonymous information available in WordNet.

Before moving to the evaluation of our model, however, it is wise to assess whether our training corpus, the BNC, actually contains the kinds of information collected in our gold standards. An easy way to do so is to count the proportion of human-generated features that co-occur with the target verbs within a given windows size, thus adapting the paradigm exploited by Schulte im Walde and Melinger (2008) to investigate verb semantic associations. Consistent with the settings of our DSMs, we fixed our window size to ± 2 sentences

and excluded from our analysis the three verbs whose absolute frequency was below the 1,000 occurrences threshold (i.e., TO TERRORIZE, TO INTERROGATE, and TO WORSHIP).

Figure 6.2 shows the proportion of verb-feature pairs from the role-based dataset (*solid line*) and from the expanded role-based dataset (*dotted line*) that co-occur in the BNC with a minimum frequency of 1, 5, 10, and 20 (x-axis). Focusing on the most appropriate threshold given the corpus size, i.e., a minimum frequency of 5, we can see that a bit more than half of the verb-feature pairs (55.56%) from the extended role-based dataset can be traced in the BNC, and this proportion decreases to less than one third if we look for the verb-feature pairs from the role-based dataset (29.18%). These numbers seem to confirm the shared belief that there are crucial differences in the information that can be extracted from corpora and the information extracted from human-elicited descriptions. Whereas some authors see corpora-derived measures as “a form of crowd-based measures, where the crowd consists of writers freely creating text on different topics” (Keuleers and Balota, 2015, p. 463), others stress the fact that corpora seem to lack many of the nonlinguistic mental properties available in the norms collections (De Deyne et al., 2015) or the fact that norms tend to represent distinctive properties of concepts, whereas texts in corpora report properties that are relevant for their communicative purposes (McRae et al., 2005b).

What is crucial for the present work, however, is the awareness that our models should not try to reach the maximum recall, but rather focus on precision. That is, a model’s performance depends on its ability to associate each verb-specific role with features that are attested in our gold standards, notwithstanding its ability to extract *all* the information available in our datasets. This is reminiscent of what happens in many Information Retrieval studies, particularly those involving web search (Manning et al., 2008), which measure the precision of the top k retrieved results. Similarly, we derive the “Precision at k ” metric by counting how many features, for each verb-specific role r_v , are attested in the gold standard.

However, the gold standard features are not directly comparable with the contextual constructions `relation-1:verb` extracted by our DSMs. We therefore simplified the latter by removing the specification of the inverse syntactic relation. This way, we are not able to distinguish constructions such as `subj-1:imprison` (*s/he imprisons*) and `obj-1:imprison` (*s/he is imprisoned*). Both contextual constructions are thus conflated into one feature: `imprison`.

To determine whether both the full model and the quasi-Davidsonian model performed better than chance, we implemented a random baseline for each model by replacing every feature with a common noun, verb, adjective, or adverb in the same frequency range. For expediency, we won’t report here the

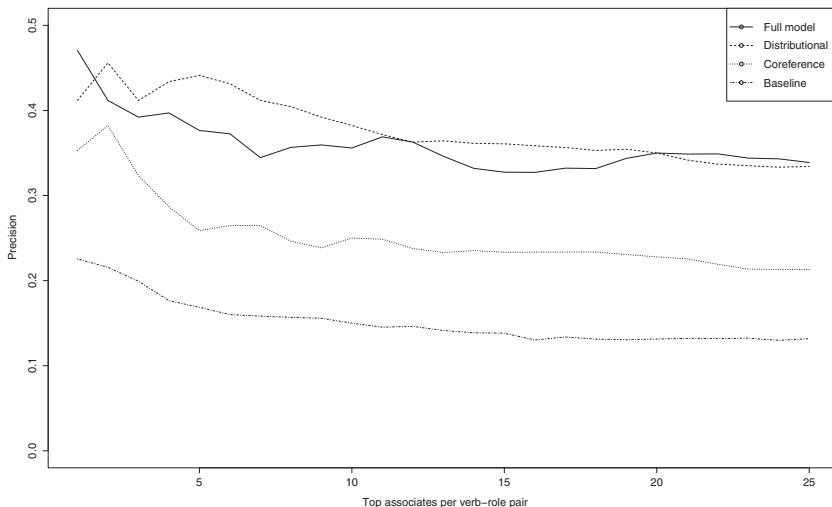


Figure 6.3 Precision of the different models evaluated against the dataset of extended role-based features.

precision of each randomized model, but we do average between them and refer to this baseline as the one obtained from the random model.

Figure 6.3 shows the precision values of the different models for different top k -selected features per verb-specific role (x-axis), evaluated against the extended role-based features. Exact values for reference values of k are reported in Table 6.4 in Section Appendix 2. Results appear to be higher than those reported by the literature on the automatic extraction of feature-like descriptions of concrete concepts (Baroni et al., 2008; Baroni and Lenci, 2010; Kelly et al., 2013), but their magnitude should be better interpreted as another confirmation of the difficulty of the task.

The best-performing models are the full model (*solid line*) and the distributional model (*dashed line*), both performing better than the coreference-based one (*dotted line*). All DSMs, moreover, performed better than the chance level (*dash-dotted line*), whose precision is around 0.15.

A similar pattern, although with lower precision scores, is obtained by evaluating the models against the role-based features, as shown by Figure 6.4 (see scores on Table 6.5). Again, all models perform better than the random baseline (precision ≈ 0.04). Again, the full model and the distributional model registered better scores than the coreference-based model.

There are, however, two main reasons why we would not take this as strong evidence against the utility of coreference-based information in modeling semantic role inferences. First of all, given the preliminary nature of our work,

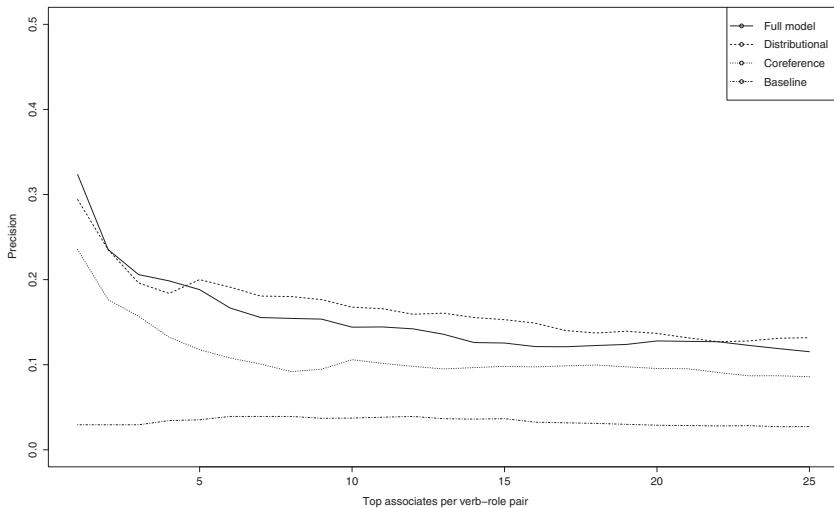


Figure 6.4 Precision of the different models evaluated against the dataset of role-based features.

we did not experiment with many of the settings that could influence the performance of both the full model and the coreference model, including the size of the context. Moreover, there is probably a joint effect of the general difficulty of the coreference annotation task (Recasens et al., 2010; Pradhan et al., 2012) and of data sparsity, due to both corpus size and the neglect of dialogue-related phenomena like implicit arguments (e.g., Ruppenhofer et al., 2011; Roth and Frank, 2013). On the other hand, the performance of the full model itself is a clue in favor of our caution. This model is basically a coreference DSM that exploits the distributional information merely to filter out unwanted features. As a consequence, it is fairly possible that the gap between the full model and the coreference DSM is due to noise that can be eliminated by using a wider context for the coreference chains, by exploiting a larger corpus, or by manually checking the relevant coreference data.

Taken together, these results appear encouraging to us, especially in light of the several limitations in the implementation of our models. Clearly, these weak spots leave plenty of room for future improvements. First of all, we overtly decided to ignore all the properties that can be inferred from dependencies headed by the fillers (in fact we used only inverse dependencies) or from superficial patterns. This will require an in-depth evaluation of the possible strategies to extract this additional information and to integrate it in our model. Moreover, it is possible that the settings we chose for our hyperparameters (e.g., number of

top fillers, association measure) are not optimal for our task. As far as the distributional space is concerned, we drew from previous experience and available comparative works (e.g., Bullinaria and Levy, 2007, 2012; Lapesa and Evert, 2014). The situation has been quite different for the use of coreference information. There was no available comparative literature, and we made our choices mainly drawing from intuition and qualitative analysis of several rounds of preliminary testing.

Finally, it is well known that the evaluation methods we chose underestimate precision. The exemplar contextual constructions in Table 6.3 from Appendix 1 illustrate this point. In this table, the constructions whose fillers are associated with the target verb-specific role in the gold standard are marked in the *match* column: two check marks for role-filler pairings that are attested in both gold standards, one check mark for those pairings that are attested only in the extended role-based norms. Even a quick look at the unmarked features associated with the verb TO ARREST shows a high number of false negatives: *s/he who arrests may even release, s/he who arrests may detain; s/he who is arrested may bail; s/he who is arrested may be oppressed; s/he who is arrested may have been recaptured; s/he who is arrested may be proclaimed* (e.g., innocent); *s/he who is arrested may be inhibited; s/he who is arrested may have abducted someone*. Arguably, we could have chosen a less conservative evaluation method, such as a feature verification paradigm. Scholars working on the automatic extraction of concrete concepts features report increases in precision as high as 0.4 when switching from a norm-based evaluation to an evaluation based on speakers' judgments (Kelly et al., 2013). Crowdsourcing techniques analogous with those developed by Reisinger et al. (2015) easily can be adapted for our purposes. However, such a choice would have come at the price of a higher number of false positives, mainly because it is often possible to find a context in which a role-feature pair may be true, even if this association is not particularly meaningful. Once again, we opted for the conservative choice, thus leaving the use of different evaluation techniques to future investigations.

In closing, it is worthwhile to stress that another consequence of the preliminary nature of our work has been the choice to restrict our target verbs to those investigated by McRae et al. (1997b). It is our opinion that the generalization of our results to other verbs would require control of many random and fixed effects, including several shades of ambiguity (e.g., lexical ambiguity, syntactic ambiguity), sociolinguistic issues (e.g., corpus data could be biased toward less prototypical uses of a verb), even theoretical considerations (some classes of verbs [e.g., light verbs] are probably harder to characterize, automatically or manually). Owing to space limitations, however, we must leave investigation of this crucial issue to future works.

6.5 Conclusion

This paper has introduced a novel unsupervised method to characterize the semantic content of verb-specific agent and patient proto-roles as bundles of presuppositions and entailment relations. Our primary intent was to test whether and to what extent semantic knowledge automatically extracted from text can be used to infer the kinds of entailments on which semantic roles are grounded. At the same time, by tackling this issue we implicitly provided evidence in favor of the idea that at least part of the knowledge about events manifests itself in the way verbs are used in a communicative environment, and that part of this generalized knowledge can be distilled from the linguistic productions available in corpus. In the view adopted in these pages, which we borrowed from Dowty (1991) and McRae et al. (1997b), it is exactly this kind of knowledge that works as a source from which the semantic content of thematic roles, by a sort of clustering process, is carved.

We evaluated different implementations of our method against a dataset of human-elicited descriptions collected with a modified version of the McRae paradigm (McRae et al., 1997b) and expanded with lexical knowledge from WordNet. In each setting, all of our models performed well above the chance level. The best-performing models were a purely syntax-based DSM and a coreference-based DSM enhanced by a syntax-based representation, both achieving a precision score between 0.35 and 0.45. Both the behavioral data and the automatically extracted verb-specific properties are freely available for downloading at <http://colinglab.humnet.unipi.it/resources/>.

The main contribution of our work, however, is not the model itself, but the demonstration that state-of-the-art computational techniques can be easily adapted to reach a decompositional description of the semantic content of thematic roles. To the best of our knowledge, the only related work in the computational linguistics literature is the one by Reisinger et al. (2015). As a consequence, we cannot but speculate over the potential applications that can benefit from our shift in perspective. However, one specific branch of research pops up immediately, i.e., the one focusing on the extraction and representation of Semantic Roles. No decompositional approach available today has the maturity to be used as a complete and usable theoretical framework, and that's probably why we're still stacked with the traditional *I-can't-define-it-but-I-know-it-when-I-see-it* stance on thematic roles, using Dowty's words (Dowty, 1989). However, the theoretical perplexities that drove the theoretical linguists to treat the atomistic view of semantic roles as an inadequate representation of the reality are strictly related to the difficulties that all researchers deal with when working with thematic roles: What is their inventory? How can they be identified? On the basis of which properties? How are they realized in the syntactic structure? The model we proposed in these pages should be seen as an attempt to

look at all these theoretical and practical issue from a different, decompositional, perspective.

Acknowledgments

The authors thank Gaia Bonucelli for taking care of the normalization phase reported in Section 6.2.2. This research received financial support from the CombiNet project (PRIN 2010-2011: *Word Combinations in Italian: theoretical and descriptive analysis, computational models, lexicographic layout and creation of a dictionary*, grant n. 20105B3HE8) funded by the Italian Ministry of Education, University and Research (MIUR).

Appendix

Appendix 1 Exemplar Features for the Verbs “to arrest” and “to punish”

Table 6.3 *Top 10 associated features per role extracted with the full model*

TO ARREST			TO PUNISH		
Role	Feature	Match ^a	Role	Feature	Match
agent	sbj-1:hold	✓	agent	sbj-1:reward	
agent	sbj-1:release		agent	sbj-1:forgive	
agent	sbj-1:charge	✓	agent	sbj-1:catch	✓
agent	sbj-1:say	✓	agent	sbj-1:deserve	
agent	sbj-1:imprison	✓✓	agent	sbj-1:doom	
agent	sbj-1:detain		agent	sbj-1:condemn	
agent	sbj-1:sentence		agent	sbj-1:compound	
agent	sbj-1:remand	✓	agent	sbj-1:tolerate	✓
agent	sbj-1:live	✓	agent	sbj-1:forfeit	
agent	sbj-1:fall		agent	sbj-1:deter	
patient	sbj-1:intern		patient	obj-1:reward	
patient	sbj-1:bail		patient	obj-1:torture	✓
patient	obj-1:oppress		patient	obj-1:lock	
patient	obj-1:recapture		patient	obj-1:unnerve	✓✓
patient	sbj-1:defy	✓	patient	obj-1:whip	
patient	sbj-1:confine	✓	patient	obj-1:humiliate	✓✓
patient	obj-1:proclaim		patient	obj-1:indulge	
patient	obj-1:inhibit		patient	obj-1:torment	✓✓
patient	sbj-1:abduct		patient	sbj-1:desperate	
patient	sbj-1:caution		patient	obj-1:elevate	

^aMatch: whether the triple *(verb, role, feature lemma)* is present in the role-based norms (✓✓), in the expanded-role-based norms (✓) or in none of the gold-standard datasets (empty cell).

Appendix 2 Precision at k of the Different Models Evaluated against the Feature-based Gold Standards

Table 6.4 Evaluation against the dataset of extended role-based features

<i>k</i>	Full model	Distributional	Coreference	Baseline
5	0.38	0.44	0.26	0.17
10	0.36	0.38	0.25	0.15
15	0.33	0.36	0.23	0.14
20	0.35	0.35	0.23	0.13
25	0.34	0.33	0.21	0.13

Table 6.5 Evaluation against the dataset of role-based features

<i>k</i>	Full model	Distributional	Coreference	Baseline
5	0.19	0.20	0.12	0.04
10	0.14	0.17	0.11	0.04
15	0.13	0.15	0.10	0.04
20	0.13	0.14	0.10	0.03
25	0.12	0.13	0.09	0.03

References

- Almuhamad, Abdulrahman, and Poesio, Massimo. 2004. Attribute-Based and Value-Based Clustering: An Evaluation. Pages 158–165 of: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.
- Almuhamad, Abdulrahman, and Poesio, Massimo. 2005. Concept Learning and Categorization from the Web. Pages 103–108 of: *Proceedings of the Annual Meeting of the Cognitive Science Society*.
- Altmann, Gerry T.M., and Kamide, Yuki. 1999. Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, **73**(3), 247–64.
- Altmann, Gerry T.M., and Kamide, Yuki. 2007. The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language*, **57**(4), 502–518.
- Andrews, Mark, Vigliocco, Gabriella, and Vinson, David P. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, **116**, 463–498.
- Ashcraft, Mark H. 1978. Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, **6**, 227–232.

- Aston, Guy, and Burnard, Lou. 1998. *The BNC handbook*. Edinburgh, UK: Edinburgh University Press.
- Baker, Collin F., Fillmore, Charles J., and Lowe, John B. 1998. The Berkeley FrameNet Project. Pages 86–90 of: *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.
- Balasubramanian, Niranjan, Soderland, Stephen, Mausam, and Etzioni, Oren. 2013. Generating Coherent Event Schemas at Scale. Pages 1721–1731 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.
- Barbu, Eduard. 2008. Combining methods to learn feature-norm-like concept descriptions. Pages 9–16 of: *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI 2008 Workshop on Distributional Semantics*.
- Barbu, Eduard, and Poesio, Massimo. 2008. A Comparison of WordNet and Feature Norms. Pages 56–73 of: *Proceedings of the 4th Global Wordnet Conference (GWC 2008)*.
- Baroni, Marco, and Lenci, Alessandro. 2010. Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, **36**(4), 673–721.
- Baroni, Marco, Evert, Stefan, and Lenci, Alessandro (eds). 2008. *Bridging the Gap between Semantic Theory and Computational Simulations: Proceedings of the ESSLLI 2008 Workshop on Distributional Semantics*.
- Baroni, Marco, Murphy, Brian, Barbu, Eduard, and Poesio, Massimo. 2010. Strudel: A Corpus-Based Semantic Model Based on Properties and Types. *Cognitive Science*, **34**, 222–254.
- Bicknell, Klinton, Elman, Jeffrey L., Hare, Mary, McRae, Ken, and Kutas, Marta. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, **63**(4), 489–505.
- Bullinaria, John A., and Levy, Joseph P. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, **39**(3), 510–526.
- Bullinaria, John A., and Levy, Joseph P. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, **44**(3), 890–907.
- Chambers, Nathanael. 2013. Event Schema Induction with a Probabilistic Entity-Driven Model. Pages 1797–1807 of: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.
- Chambers, Nathanael, and Jurafsky, Daniel. 2008. Unsupervised Learning of Narrative Event Chains. Pages 789–797 of: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*.
- Chambers, Nathanael, and Jurafsky, Daniel. 2009. Unsupervised Learning of Narrative Schemas and Their Participants. Pages 602–610 of: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.
- Cheung, Jackie Chi Kit, Poon, Hoifung, and Vanderwende, Lucy. 2013. Probabilistic frame induction. Pages 837–846 of: *Proceedings of the 2013 Conference of the*

- North American Chapter of the Association for Computational Linguistics: Human Language Technologies.*
- Church, Kenneth Ward, and Hanks, Patrick. 1991. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, **16**(1), 22–29.
- Collins, Allan M., and Loftus, Elizabeth F. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, **82**, 407–428.
- Collobert, Ronan, Weston, Jason, Bottou, Léon, Karlen, Michael, Kuksa, Pavel, and Kavukcuoglu, Koray. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, **12**, 2493–2537.
- Cree, George S., McRae, Ken, and McNorgan, Chris. 1999. An attractor model of lexical conceptual processing: simulating semantic priming. *Cognitive Science*, **23**(3), 371–414.
- De Deyne, Simon, Verheyen, Steven, Ameel, Eef, Vanpaemel, Wolf, Dry, Matthew J., Voorspoels, Wouter, and Storms, Gert. 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, **40**(4), 1030–48.
- De Deyne, Simon, Verheyen, Steven, and Storms, Gert. 2015. The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The Quarterly Journal of Experimental Psychology*, **68**(8), 1643–1664.
- Devereux, Barry J., Pilkington, Nicholas, Poibeau, Thierry, and Korhonen, Anna. 2009. Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data. *Research on Language and Computation*, **7**(2-4), 137–170.
- Devereux, Barry J., Tyler, Lorraine K., Geertzen, Jeroen, and Randall, Billi. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, **46**(5), 1119–1127.
- Dowty, David. 1989. On the Semantic Content of the Notion of Thematic Role. Pages 69–130 of: Chierchia, Gennaro, Partee, Barbara H., and Turner, Raymond (eds), *Properties Types and Meaning*, Vol. II, Semantic Issues. Dordrecht; Kluwer Academic Publishers.
- Dowty, David. 1991. Thematic Proto-Roles and Argument Selection. *Language*, **67**(3), 547–619.
- Erk, Katrin. 2007. A Simple, Similarity-based Model for Selectional Preferences. Pages 216–223 of: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.
- Erk, Katrin, Padó, Sebastian, and Padó, Ulrike. 2010. A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, **36**(4), 723–763.
- Evert, Stefan. 2009. Corpora and Collocations. Chap. 58, pages 1212–1248 of: Lüdeling, Anke, and Kytö, Merja (eds), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.
- Fagarasan, Luana, Vecchi, Eva Maria, and Clark, Stephen. 2015. From distributional semantics to feature norms: Grounding semantic models in human perceptual data. Pages 52–57 of: *Proceedings of the 11th International Conference on Computational Semantics*.

- Fellbaum, Christiane. 1998. *WordNet - An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Ferretti, Todd R., McRae, Ken, and Hatherell, Andrea. 2001. Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, **44**(4), 516–547.
- Ferretti, Todd R., Kutas, Marta, and McRae, Ken. 2007. Verb Aspect and the Activation of Event Knowledge. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **33**(1), 182–196.
- Fillmore, Charles J. 1968. The Case for Case. Pages 0–88 of: Bach, Emmon, and Harms, Robert T. (eds), *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston.
- Fort, Karën, Adda, Gilles, and Cohen, K. Bretonnel. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics*, **37**(2), 413–420.
- Frermann, Lea, Titov, Ivan, and Pinkal, Manfred. 2014. A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge. Pages 49–57 of: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.
- Garrard, Peter, Ralph, Matthew. A. Lambon, Hodges, John R., and Patterson, Karalyn. 2001. Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, **18**(2), 125–174.
- Gildea, Daniel, and Jurafsky, Daniel. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, **28**(3), 245–288.
- Greenberg, Clayton, Sayeed, Asad B., and Demberg, Vera. 2015. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. Pages 21–31 of: *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*.
- Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Dordrecht: Kluwer Academic Publishers.
- Gruber, Jeffrey S. 1965. Studies in lexical relations. PhD thesis, Massachusetts Institute of Technology.
- Hampton, James A. 1979. Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **18**(4), 441 – 461.
- Hare, Mary, Elman, Jeffrey L., Tabaczynski, Tracy, and McRae, Ken. 2009. The Wind Chilled the Spectators, but the Wine Just Chilled: Sense, Structure, and Sentence Comprehension. *Cognitive Science*, **33**(4), 610–628.
- Harris, Zellig S. 1954. Distributional structure. *Word*, **10**, 146–162.
- Hermann, Karl Moritz, Das, Dipanjan, Weston, Jason, and Ganchev, Kuzman. 2014. Semantic Frame Identification with Distributed Word Representations. Pages 1448–1458 of: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.
- Hinton, Geoffrey E., and Shallice, Tim. 1991. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, **98**, 74–95.
- Jackendoff, Ray. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.
- Jackendoff, Ray. 1987. The Status of Thematic Relations in Linguistic Theory. *Linguistic Inquiry*, **18**(3), 369–411.

- Kamide, Yuki, Altmann, Gerry T.M., and Haywood, Sarah L. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, **49**(1), 133 – 156.
- Kelly, Colin, Devereux, Barry, and Korhonen, Anna. 2010. Acquiring Human-like Feature-based Conceptual Representations from Corpora. Pages 61–69 of: *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*.
- Kelly, Colin, Devereux, Barry J., and Korhonen, Anna. 2013. Automatic Extraction of Property Norm-Like Data From Large Text Corpora. *Cognitive Science*, **38**(4), 638–682.
- Keuleers, Emmanuel, and Balota, David A. 2015. Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *The Quarterly Journal of Experimental Psychology*, **68**(8), 1457–1468.
- Kingsbury, Paul, and Palmer, Martha. 2003. PropBank: the Next Level of TreeBank. In: *Proceedings of Treebanks and Lexical Theories*.
- Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, **42**(1), 21–40.
- Kipper-Schuler, K. 2005. Verbnet: A broad-coverage, comprehensive verb lexicon. PhD thesis, University of Pennsylvania.
- Kittur, Aniket, Nickerson, Jeffrey V., Bernstein, Michael, Gerber, Elizabeth, Shaw, Aaron, Zimmerman, John, Lease, Matt, and Horton, John. 2013. The Future of Crowd Work. In: *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*.
- Kremer, Gerhard, and Baroni, Marco. 2011. A Set of Semantic Norms for German and Italian. *Behavior Research Methods*, **43**(1), 97–109.
- Lapesa, Gabriella, and Evert, Stefan. 2014. A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection. *Transactions of the Association for Computational Linguistics*, **2**, 531–545.
- Lebani, Gianluca E. 2012. STaRS.sys: designing and building a commonsense-knowledge enriched wordnet for therapeutic purposes. PhD thesis, University of Trento.
- Lebani, Gianluca E., Bondielli, Alessandro, and Lenci, Alessandro. 2015. You Are What You Do. An Empirical Characterization of the Semantic Content of the Thematic Roles for a Group of Italian Verbs. *Journal of Cognitive Science*, **16**(4), 401–430.
- Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. A foreword. *Italian Journal of Linguistics*, **20**(1), 1–30.
- Lenci, Alessandro. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. Pages 58–66 of: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*.
- Lenci, Alessandro, Baroni, Marco, Cazzolli, Giulia, and Marotta, Giovanna. 2013. BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, **45**(4), 1218–1233.
- Levin, Beth, and Rappaport Hovav, Malka. 2005. *Argument Realization*. Cambridge, UK: Cambridge University Press.
- Lin, Dekang. 1998. Automatic Retrieval and Clustering of Similar Words. Pages 768–774 of: *Proceedings of the 36th Annual Meeting of the Association for*

- Computational Linguistics and 17th International Conference on Computational Linguistics.*
- Liu, Ding, and Gildea, Daniel. 2010. Semantic Role Features for Machine Translation. Pages 716–724 of: *Proceedings of the 23rd International Conference on Computational Linguistics*.
- Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. 2008. *An Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.
- Màrquez, Lluís, Carreras, Xavier, Litkowski, Kenneth C., Stevenson, Suzanne (eds). 2008. *Computational Linguistics: Special Issue on Semantic Role Labeling*, **34**(2).
- Matsuki, Kazunaga, Chow, Tracy, Hare, Mary, Elman, Jeffrey L., Scheepers, Christoph, and McRae, Ken. 2011. Event-Based Plausibility Immediately Influences On-Line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**(4), 913–934.
- McRae, Ken, and Cree, George S. 2001. Factors underlying category-specific semantic deficits. In: Forde, E. M. E., and Humphreys, G. (eds), *Category Specificity in Mind and Brain*. Hove, East Sussex, UK: Psychology Press.
- McRae, Ken, and Matsuki, Kazunaga. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, **3**(6), 1417–1429.
- McRae, Ken, De Sa, Virginia R., and Seidenberg, Mark S. 1997a. On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, **126**(2), 99–130.
- McRae, Ken, Ferretti, Todd R., and Amyote, Liane. 1997b. Thematic Roles as Verb-specific Concepts. *Language and Cognitive Processes*, **12**(2/3), 137–176.
- McRae, Ken, Spivey-Knowlton, Michael J., and Tanenhaus, Michael K. 1998. Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, **38**(3), 283–312.
- McRae, Ken, Hare, Mary, Elman, Jeffrey L., and Ferretti, Todd R. 2005a. A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, **33**(7), 1174–1184.
- McRae, Ken, Cree, George S., Seidenberg, Mark S., and McNorgan, Chris. 2005b. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**(4), 547–59.
- Miller, George A., and Charles, Walter G. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.
- Montefinese, Maria, Ambrosini, Ettore, Fairfield, Beth, and Mammarella, Nicola. 2013. Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, **45**(2), 440–461.
- Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandrs, Marinov, Svetoslav, and Marsi, Erwin. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(6), 95–135.
- Padó, Sebastian, and Lapata, Mirella. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, **33**(2), 161–199.
- Padó, Ulrike. 2007. The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing. PhD thesis, Saarland University.

- Palmer, Martha, Gildea, Daniel, and Xue, Nianwen. 2010. Semantic Role Labeling. *Synthesis Lectures on Human Language Technologies*, 3(1), 1–103.
- Poesio, Massimo, Barbu, Eduard, Giuliano, Claudio, and Romano, Lorenza. 2008. Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In: *Proceedings of the 3rd Workshop on Ontology Learning and Population*.
- Pradhan, Sameer, Moschitti, Alessandro, Xue, Nianwen, Uryupina, Olga, and Zhang, Yuchen. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. Pages 1–40 of: *Joint Conference on EMNLP and CoNLL - Shared Task*.
- Preiss, Judita, Briscoe, Ted, and Korhonen, Anna. 2007. A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. Pages 912–919 of: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*.
- Recasens, Marta, Márquez, Lluís, Sapena, Emili, Martí, M Antònia, Taulé, Mariona, Hoste, Véronique, Poesio, Massimo, and Versley, Yannick. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. Pages 1–8 of: *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*.
- Regneri, Michaela, Koller, Alexander, and Pinkal, Manfred. 2010. Learning Script Knowledge with Web Experiments. Pages 979–988 of: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.
- Reisinger, Drew, Rudinger, Rachel, Ferraro, Francis, Harman, Craig, Rawlins, Kyle, and Van Durme, Benjamin. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3, 475–488.
- Roller, Stephen, and Schulte im Walde, Sabine. 2014. Feature Norms of German Noun Compounds. Pages 104–108 of: *Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014)*.
- Rosch, Eleanor, and Mervis, Carolyn B. 1975. Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, 573–605.
- Roth, Michael, and Frank, Anette. 2013. Automatically Identifying Implicit Arguments to Improve Argument Linking and Coherence Modeling. Pages 306–316 of: *Second Joint Conference on Lexical and Computational Semantics (*SEM)*.
- Roth, Michael, and Lapata, Mirella. 2015. Context-aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3, 449–460.
- Ruppenhofer, Josef, Gorinski, Philip, and Sporleder, Caroline. 2011. In Search of Missing Arguments: A Linguistic Approach. Pages 331–338 of: *Proceedings of Recent Advances in Natural Language Processing*.
- Sahlgren, Magnus. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in highdimensional vector spaces. PhD thesis, Stockholm University.
- Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1), 33–53.
- Sartori, Giuseppe, and Lombardi, Luigi. 2004. Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, 16(3), 439–52.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. Pages 44–49 of: *Proceedings of the international conference on new methods in language processing*.

- Schulte im Walde, Sabine, and Melinger, Alissa. 2008. An in-depth look into the co-occurrence distribution of semantic associates. *Italian Journal of Linguistics*, **20**(1), 87–123.
- Shen, Dan, and Lapata, Mirella. 2007. Using Semantic Roles to Improve Question Answering. Pages 12–21 of: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.
- Smith, Edward E., Shoben, Edward J., and Rips, Lance J. 1974. Structure and Process in Semantic Memory: A Featural Model for Semantic Decisions. *Psychological Review*, **81**(3), 18–47.
- Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, and Ng, Andrew Y. 2008. Cheap and Fast-but is It Good? Evaluating Non-expert Annotations for Natural Language Tasks. Pages 254–263 of: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Steyvers, Mark, Smyth, Padhraic, and Chemuduganta, Chaitanya. 2011. Combining Background Knowledge and Learned Topics. *Topics in Cognitive Science*, **3**(1), 18–47.
- Storms, Gert, Navarro, Daniel J., and Lee, Michael D. (eds). 2010. *Acta Psychologica Special Issue on Formal Modeling of Semantic Concepts*. Vol. 133 (3).
- Taylor, Wilson. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism Quarterly*, **30**, 415–433.
- Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.
- Traxler, Matthew J., Foss, Donald J., Seely, Rachel E., Kaup, Barbara, and Morris, Robin K. 2001. Priming in Sentence Processing: Intralexical Spreading Activation, Schemas, and Situation Models. *Journal of Psycholinguistic Research*, **29**(6), 581–595.
- Turney, Peter D., and Pantel, Patrick. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, **37**, 141–188.
- Van Valin, Robert D. Jr. 1999. Generalized semantic roles and the syntax-semantics interface. Pages 373–389 of: Corblin, F., Dobrovie-Sorin, C., and Marandin, J.-M. (eds), *Empirical issues in formal syntax and semantics*. The Hague: Thesus.
- Versley, Yannick, Ponzetto, Simone Paolo, Poesio, Massimo, Eidelman, Vladimir, Jern, Alan, Smith, Jason, Yang, Xiaofeng, and Moschitti, Alessandro. 2008. BART: A Modular Toolkit for Coreference Resolution. Pages 9–12 of: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*.
- Vigliocco, Gabriella, Vinson, David P., Lewis, William D., and Garrett, Merrill F. 2004. Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, **48**(4), 422–88.
- Vigliocco, Gabriella, Warren, Jane, Siri, Simona, Arciuli, Joanne, Scott, Sophie, and Wise, Richard. 2006. The role of semantics and grammatical class in the neural representation of words. *Cerebral Cortex*, **16**(12), 1790–1796.
- Vinson, David P., and Vigliocco, Gabriella. 2008. Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, **40**, 183–190.
- Vinson, David P., Vigliocco, Gabriella, Cappa, Stefano F., and Siri, Simona. 2003. The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, **86**, 347–365.

- Wu, Shumin, and Palmer, Martha. 2011. Semantic Mapping Using Automatic Word Alignment and Semantic Role Labeling. Pages 21–30 of: *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*.
- Zapirain, Benat, Agirre, Eneko, Márquez, Lluís, and Surdeanu, Mihai. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, **39**(3), 631–663.

7 Native Language Identification on EFCAMDAT

*Xiao Jiang, Yan Huang, Yufan Guo, Jeroen Geertzen,
Theodora Alexopoulou, Lin Sun, and Anna Korhonen*

Abstract

Native Language Identification (NLI) is a task aimed at determining the native language (L1) of learners of second language (L2) on the basis of their written texts. To date, research on NLI has focused on relatively small corpora. We apply NLI to EFCAMDAT, an L2 English learner corpus that is not only multiple times larger than previous L2 corpora but also provides pseudo-longitudinal data across several proficiency levels. Based on accurate machine learning with a wide range of linguistic features, our investigation reveals interesting patterns in the longitudinal data that are useful for both further development of NLI and its application to research on L2 acquisition.

7.1 Introduction

Native language identification (NLI) is a task aimed at detecting the native language (L1) of writers on the basis of their second language (L2) production. NLI is important for natural language processing (NLP) applications including language tutoring systems and authorship profiling. Moreover, NLI can offer useful empirical data for research on L2 acquisition. For example, NLI can shed light on how L1 background influences L2 learning, and on differences between the writings of L2 learners across different L1 backgrounds.

To date, studies on NLI have focused on relatively small learner corpora. Furthermore, none of them have investigated the influence of L1s across L2 proficiency levels. Our work takes the first step toward addressing these problems. We apply NLI to EFCAMDAT, the EF-Cambridge Open Language Database (Geertzen, Alexopoulou, and Korhonen, 2013),¹ an open-access L2 learner corpus.

EFCAMDAT consists of writings of learners submitted to *Englishtown*, the online school of EF. EFCAMDAT stands out for its size, diversity of student

¹ <http://corpus.mml.cam.ac.uk/efcamdat>

backgrounds, and coverage of the proficiency levels. The first release of 2013 (Geertzen, Alexopoulou, and Korhonen, 2013), on which this paper is based, amounts to 30 million words, a corpus multiple times larger than any other available L2 corpora. Using a standard machine learning-based methodology for NLI, we explore the optimal linguistic features for NLI on this data at different proficiency levels. We discover interesting patterns that can be useful for both further development of NLI and its application to research on L2 acquisition.

In this introductory section, we first review the history of research on NLI, and introduce the data sets that have been used in earlier NLI research. We then summarise our contribution briefly. Section 7.2 describes our data set EFCAM-DAT in detail, Section 7.3 describes our research method, Section 7.4 presents our empirical results and qualitative analysis, and finally Section 7.5 presents our conclusions.

7.1.1 Prior studies on NLI

The first study on NLI was conducted by Tomokiyo and Jones (2001). While their original goal was to develop techniques for detecting nonnative speech, they actually built a Naive Bayes classifier to distinguish between the native speakers of Chinese and Japanese according to the transcripts of their English utterances. Based on word n-grams in which nouns were replaced by their part-of-speech (POS) tags, the classifier achieved a remarkable accuracy of 100%. However, the generalizability of the result is questionable considering the limited population of the subjects: only eight English and six Chinese speakers were involved.

Koppel, Schler, and Zigdon (2005) regarded NLI as a subtask of authorship attribution, and constructed a Support Vector Machine (SVM) classifier for NLI on five native language backgrounds. With 1,035 features, including function words, character n-grams, error types extracted by the grammar checker of Microsoft Word, rare words, and POS n-grams, the classifier achieved an accuracy of 80.2%. The study became a benchmark for most subsequent research.

Tsur and Rappoport (2007) replicated the study of Koppel, Schler, and Zigdon (2005), but sampled the texts differently and fed only one type of features to the classifier at a time. They found that character bi-grams were most discriminatory for NLI; the highest classification accuracy was 66%. The accuracy of bi-gram classifier remained high even when the dominant content words or function words were removed from the texts. It was concluded that the word choice of second-language writing might be subject to the phonology of writers' native languages.

Wong and Dras (2009) also followed the study of Koppel, Schler, and Zigdon (2005). They extended the number of native language backgrounds in

NLI to seven. In addition to the feature sets of Koppel, Schler, and Zigdon (2005), they studied the discriminatory power of three types of syntactic errors extracted by an automatic grammatical checker. However, results showed that the effect of these errors was not prominent, which might be attributed to the high false-positive rate of the grammatical checker, as well as the limited types of the errors. The highest classification accuracy with combined features was 73.71%. Nevertheless, the value of learner errors for NLI was highlighted by Kochmar (2011) later. He demonstrated that the error types which were typical in the English writings by native speakers of various Indo-European languages were useful for pairwise classification of these native language backgrounds. The accuracy of error-based classifiers ranges from 47.92% to 59.98%.

In a subsequent study, Wong and Dras (2011) introduced two types of syntactic features to NLI: one consists of production rules extracted from Context Free Grammar (CFG) parse tree, while the other contains reranking features previously used to select the best tree derivatives returned by parsers (Charniak and Johnson, 2005). These features led to a remarkable reduction of 30% in classification errors over the baseline. The highest classification accuracy was 81.71%. Based on the same data, Wong, Dras, and Johnson (2011; 2012) continued to investigate novel features, which included topical features extracted by Latent Dirichlet Allocation (LDA) and mixed POS and function word n-grams sifted by adaptor grammar; the latter were found to be useful in NLI. Combined with mixed n-grams containing POS and function words, it achieved a classification accuracy of 75.71%.

A number of studies were then conducted following the same data setting of Wong and Dras (2011). Ahn (2011) found that the effect of character tri-grams was highly correlated with that of word uni-grams, and that their contribution to the NLI classification accuracy was attributable to topic biases. Bykh and Meurers (2012) studied how the effect of word and POS n-grams on NLI changed with their length. They accomplished a classification accuracy of 89.71% using recurring n-grams, which were n-grams that appeared in more than two texts in sets. Swanson and Charniak (2012) explored the effect of a novel feature type, Tree Substitution Grammar (TSG) fragments, and found that the classifiers built on such features outperformed those built on CFG production rules (78.4% versus 72.6% in accuracy).

Jarvis and Crossley (2012) conducted a series of research into NLI aiming at detecting L1 transfer. With Latent Discriminant Analysis (LDA), they systematically investigated both the independent and combined effects of word n-grams, Coh-Metrix stylistic features (Graesser et al., 2004), and manually annotated errors in NLI. Their results also justified the effect of errors: when different feature types were fed independently to the classifier, the fine-grained manually annotated errors led to the highest three-way classification accuracy of 65.5%.

Tetreault et al. (2012) trained classifiers for separate feature types, and combined their outputs to build an ensemble classifier that produced the highest score so far for seven-way NLI (90.1%) on the commonly used International Corpus of Learner English (ICLE) (Granger, 2003). The ICLE data used by Tetreault et al. (2012) underwent some corrective processing, and were controlled for topic distribution, so they were not entirely the same as those used in earlier studies. They also conducted cross-corpus evaluation, and found that the classifiers trained on the small ICLE data cannot generalize well to larger corpora, while those trained on the larger corpora can generalize to the ICLE data.

As a response to an increased research interest in NLI, the first NLI shared task was held in 2013 (Tetreault, Blanchard, and Cahill, 2013), attracting 29 participating teams. A number of novel features were tried, such as the round-trip translation of English words which were intended to capture lexical preferences of each native language group (Lavergne et al., 2013), brown clusters that were produced by hierarchical clustering of words based on the context, and restored tags that were achieved by removing some words from text and recovering the words according to an n-gram language model to capture consistent omission or misuse of these words (Tsvetkov et al., 2013). However, most of these new features brought marginal or no empirical improvement to NLI. The most successful teams (Jarvis, Bestgen, and Pepper, 2013) generally involved long n-grams of words (at least quad-grams) and characters (up to 9-grams) in building their classifiers. The system that ranked third in the closed NLI shared task was solely built on characters using Kernel Ridge Regression (KRR) (Popescu and Ionescu, 2013). By adding an intersection kernel in a follow-up study, the same team achieved an accuracy that is 1.7% higher than the top scoring team in the 2013 shared task (Ionescu, Popescu, and Cahill, 2014).

While the aforementioned studies only focused on English L2 data, Malmasi and Dras extended the study of NLI to Arabic L2 (Malmasi and Dras, 2014a) and Chinese L2 (Malmasi and Dras, 2014b). They achieved accuracies of 41.0% and 70.61% on these languages, respectively, using function words, context-free grammar production rules, and POS n-grams as features. Furthermore, the authors found the classification accuracies on Chinese and English L2 data were similar.

There has also been research into comparing different classifiers for NLI. The results have been mixed. During an author-profiling study in which the native language was also one of the dependent variables, Estival et al. (2007) compared a number of classifiers including decision trees, SVM and ensemble learning, etc., and found that the Random Forest (RF) classifier with feature selection based on information gained best results for NLI on an e-mail data set. Jarvis (2011) studied 20 classifiers on NLI and found LDA to be the most

accurate, while RF and commonly used SVM classifiers were substantially worse. However, it should be noted that their study only used word n-grams as features, and that the best system in the 2013 shared task employed SVM (Jarvis, Bestgen, and Pepper, 2013).

7.1.2 *Prior Data Sets*

In terms of data sets, most previous studies on NLI employed the International Corpus of Learner English (ICLE) corpus (Granger, 2003), developed at the Centre for English Corpus Linguistics at the University of Louvain, Belgium.² The ICLE corpus was specifically designed for the study of English writings from nonnative speakers. It includes English writings from university students worldwide. These students were roughly at the higher-intermediate or advance English proficiency level. The corpus contains several subcorpora, each of which contains writings of students of one of the following native languages: Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish. In total, there are 3,640 writings and 2.5 million words in the corpus. Most writings are argumentative essays or literature examination papers. Example topics are “Crime does not pay” and “The role of censorship in Western society.” Each writings must be at least 500 words long. On average there are 690 words per writing.

While ICLE has been widely used in NLI studies (Koppel, Schler, and Zgidon, 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009; Wong and Dras, 2011; Ahn, 2011; Bykh and Meurers, 2012; Jarvis and Crossley, 2012), it is subject to topic biases (Ahn, 2011; Brooke and Hirst, 2011; Tetreault et al., 2012). Since the topics for the essays in ICLE were chosen individually by each university involved in the project, some topics are relevant to students only from a specific native language background. For example, many of the common topics in the French subcorpus concern the relatively esoteric subjects of literature, religion, and politics, while most of the Japanese subcorpus consists of more personal topics, ranging from experience as an English learner to one’s favorite travel destination. Thus, it is inevitable that the writings by French authors appear to be more formal while those by Japanese authors appear to be more narrative and colloquial. In this case, the classification accuracy of native language identification is conflated by that of topic classification. Brooke and Hirst (2011) demonstrated the topic biases in ICLE through cross-corpus evaluation: they trained a classifier on ICLE data, and applied it to a new corpus Lang-8³ that contains 22-million-word short journal entries covering a range of topics; the classification accuracy dropped by more than 54%. Brooke and Hirst (2011) believed that the topic biases in ICLE were the cause of the sharp drop, though

² <http://www.uclouvain.be/en-317607.html>

³ <http://lang-8.com>

the drop may also have been aggravated by the heterogeneous and incoherent nature of Lang-8 (Bykh and Meurers, 2012). Meanwhile, there is other empirical evidence of the topic biases in ICLE: Ahn (2011) demonstrated that when excluding topic words unique to each native language subcorpus, the performance of classifier based on uni-grams deteriorated dramatically.

Also, the distribution of L2 proficiency in ICLE was found to be unbalanced across different native language groups, which could also conflate the classification accuracy of NLI. Bestgen, Granger, Thewissen, et al. (2012) rated their data from ICLE in the Common European Framework (CEF), and found that there was significant difference in the English proficiency level of the French, German, and Spanish groups, and that the occurrence of learner errors was negatively correlated with the L2 proficiency level of the writers. For research that is intended to analyze L1 transfer, the effect of L2 proficiency level has to be controlled.

Some prior NLI studies have also used other data sets. Al-Rfou (2012) exploited English Wikipedia comments in their research. As 47% of the 60,000 Wikipedia editors had specified their native languages as non-English, the editor comments could be used for NLI. Al-Rfou (2012) built a corpus that contained more than 12 million words from 9,857 distinct authors representing the top 20 most frequently used native languages. They argued that working on the Wikipedia data set was more challenging because each individual writing was shorter, and that the writings were diverse in topics as well as the English proficiency levels of their writers. Kochmar (2011) used Cambridge Learner Corpus, which consists of more than 200,000 writings produced during Cambridge ESOL English exams by students from 217 countries. As stated earlier, the corpus has manual error tags, which enabled Kochmar (2011) to study the usefulness of error-driven features in NLI. Estival et al. (2007) used 9,836 English-language e-mails written by English, Arabic, and Spanish people. Jarvis and Crossley (2012) utilized narrative film descriptions on only one topic for studying the effect of uni-grams. Tetreault et al. (2012) adopted TOEFL 11 corpus, which contained 11,000 essays written by test-takers during The Test of English as a Foreign Language (TOEFL), with 1,000 texts for each of 11 native language groups. The corpus was then enlarged to 1,100 texts for each group (Blanchard et al., 2013), and was used in the closed NLI shared task in 2013 (Tetreault, Blanchard, and Cahill, 2013).

7.1.3 Our contribution

In this study we employ the recently released EFCAMDAT corpus for NLI. Since the data come from a live educational context, where the writings are submitted by a large number of students from diverse backgrounds, the corpus provides a rich resource for the development and evaluation of NLI as well as for linguistic studies. We explore the potential and challenges of NLI when applied

to this new longitudinal data set. We report experiments where a rich set of linguistic features were extracted from this corpus (including word and character n-grams, POS n-grams, production rules, and grammatical relations) using state-of-the-art NLP and classified using Support Vector Machines (SVM). We conduct a quantitative and qualitative evaluation of the performance of different features and compare, for the first time, the performance of different features at different proficiency levels. We observe patterns interesting for the development of both NLI and L2 acquisition research: the top performing features differ across proficiency levels and such patterns can have relevance for research on L2 acquisition.

7.2 Data

EFCAMDAT was developed at the University of Cambridge, in collaboration with Education First (EF) – an international organization of teaching English as a foreign language. The corpus consists of writings submitted to EF English-town,⁴ the online school of EF. The EF curriculum is organised along 16 teaching levels, which are aligned to the Common European Framework of Reference Levels (CEFR). EF teaching levels 1–3 correspond to CEFR A1, 4–6 to CEFR A2, 7–9 to CEFR B1, 10–12 to CEFR B2 and 13–16 to C1. Each teaching level consists of eight lessons; at the end of each lesson there is a writing assignment that learners submit for correction by (human) EF teachers. When learners sign up for a course, they take a placement test to identify their initial proficiency level. The majority of learners in EFCAMDAT complete one to three teaching levels. For further details, see Geertzen, Alexopoulou, and Korhonen (2013).

Each writing is accompanied with metadata including its submission date, EF teaching level, teaching unit and lesson title, topic/task ID, as well as a grade marked by a human grader from EF; metadata for authors include an anonymous id, country, and nationality. A combination of nationality and country have been used as a proxy to the native language background.

Table 7.1 shows the current total number of documents, words, learners, nationalities, and proficiency levels covered by EFCAMDAT (as of October 2013).

Table 7.2 shows the top 10 nationalities with most writings, from 175 different nationalities. All nationalities in the Top 10 list have their unique native languages. The native language of Brazilians is Portuguese, and Spanish is that of Mexicans. Writings from these 10 nationalities make up 90% of the whole corpus.

⁴ <http://www.englishtown.com/>

Table 7.1 *The statistics of EFCAMDAT*

	Count
Documents	423,373
Words	30,763,521
Learners	76,002
Nationalities	175
Proficiency levels	16

Table 7.2 *Top 10 nationalities by the number of writings*

Rank	Nationality	# Writings	Rank	Nationality	# Writings
1	Chinese	162,256	6	Mexican	15,802
2	Brazilian	71,182	7	French	14,067
3	Russian	63,470	8	Saudi Arabians	7,743
4	Italian	19,304	9	Americans	6,712
5	German	17,030	10	Japanese	6,027

Table 7.3 shows the number of writings at each of the 16 different proficiency levels. As can be seen, most writings come from low proficiency levels. Writings from levels 1 to 3 make up 41% of the whole corpus; writings from levels 4 to 7 make up another 40% of the whole corpus; thus, almost 80% of writings fall broadly within CEFR level A.

There are 128 tasks following topics such as the ones shown in Table 7.4.

EFCAMDAT is different from the frequently used ICLE corpus in many ways. First, EFCAMDAT is an order-of-magnitude larger than ICLE. The former contains half a million writings from 76,000 authors that total 30 million words, while the latter only contains 3,640 writings totaling 2.5 million words. As new

Table 7.3 *Number of writings at 16 proficiency levels*

Lvl	# Writings	Lvl	# Writings
1	91,948	9	16,056
2	38,220	10	21,710
3	43,776	11	10,180
4	74,700	12	5,695
5	35,922	13	3,985
6	20,214	14	1,567
7	39,700	15	745
8	18,456	16	499

Table 7.4 *Example topics for 16 proficiency levels*

Lvl	Example Topics	Lvl	Example Topics
1	Giving instructions to play a game	9	Giving feedback to a restaurant
2	Writing a birthday invitation	10	Helping a friend find a job
3	Renovating your home	11	Writing a movie review
4	Describing your family in an e-mail	12	Entering a writing competition
5	Giving suggestions about clothing	13	Writing a campaign speech
6	Writing a movie plot	14	Applying for sponsorship
7	Writing a letter of complaint	15	Covering a news story
8	Making a dinner party menu	16	Criticizing a celebrity

data keep coming in, the size of EFCAMDAT continues to grow. Second, the authors in EFCAMDAT are more diverse. They come from 175 different countries and represent 16 different proficiency levels. In contrast, the authors of ICLE are college students at roughly the same age worldwide and have advanced English proficiency. The average writing length in ICLE (500–1,000 words per writing) is much longer than that in EFCAMDAT (50–120 words per writing). Third, the topics are diverse in EFCAMDAT. There are 128 different task prompts across the 16 proficiency levels involving a variety of narrative, descriptive, and argumentative tasks from everyday communication such as “Introducing yourself by e-mail” (Level 1) to more complex tasks like “writing a movie plot” (Level 6). In contrast, most writings in ICLE are argumentative essays or literature examination paper. Being a collaboration project with many international universities, ICLE granted each university the right to assign its own topics. Therefore, the topics vary across different L1 language groups. In EFCAMDAT, the topics set for the writings at certain proficiency level are the same across all L1 backgrounds.

Performing NLI on EFCAMDAT is challenging. Being a real-world data set from EF school, the huge diversity of learner backgrounds makes the corpus noisy. Also, the average length for each writing is much shorter, and we can extract less information per writing in classification experiments. Moreover, the writings of EFCAMDAT have a relatively limited vocabulary and structure. Each writing corresponds to a lesson, and the students tended to replicate the vocabulary and sentence structure they just learnt from that lesson. Therefore, many writings from the same topic may present similar vocabulary and structure, even though they come from different learners with different native language backgrounds. Such similarity makes it more difficult to distinguish writings among different native language backgrounds.

Since we wanted to investigate different proficiency levels and not all levels had sufficient data for adequate NLI performance at the time of this experiment, we merged proficiency levels into four groups as to avoid data sparsity, as shown in Table 7.5. In doing this we have largely respected the alignment of

Table 7.5 *A subset of EFCAMDAT used in this work*

Group	Documents	Words
Lvl 1–3	44,362	1,910,674
Lvl 4–7	50,593	4,223,560
Lvl 8–11	15,095	1,726,093
Lvl 12–16	2,459	359,563

EF teaching levels to CEFR: levels 1–3 correspond to CEFR A1; the grouping 4–7 broadly aligns with CEFR A2 even though it includes EF 7 which corresponds to early B1; levels 8–11 align with the intermediate CEFR level B1/B2; while 12–16 broadly align with CEFR advanced C. We focused on three major nationalities – Chinese, Brazilian, and Russian – which jointly cover 78% of the corpus and yield a reasonably large training and test sets for NLI. We excluded some essays to ensure that each group contains approximately the same amount of data.

7.3 Methods

This section describes the methodology for building our NLI system on EFCAMDAT.

7.3.1 Features

We investigated a variety of lexical and syntactic features used in previous NLI works:

Word n-gram The sequence of adjacent words with length n . For example, for sentence I speak English, the word bi-gram ($n = 2$) features are I speak and speak English.

We experimented with word n-grams of different orders ($n = 1$ to 4) under different settings. These settings included whether to convert all letters to lower case (LC), to perform stemming (STM), to remove stop words and punctuation (RMSTP), to filter out nonalphanumeric words (ALPHANUM), and to filter out n-grams that were less frequent (MF) in the whole data set, or n-grams that had low recurring frequencies across different writings (RMF), i.e., appearing in fewer than a threshold number of different writings. LC, STM, RMSTP, and ALPHANUM were supposed to prevent overly specific features that may lead to data sparsity issues. MF and RMF were supposed to discard any rare features that were less informative and were potential noises to classifiers. The function word

features mentioned in many previous literatures were treated as another special setting for word uni-gram – only function words were kept as features while any other n-grams were discarded.

Character n-gram The sequence of adjacent characters of length n. For example, for sentence I speak English, some of the character tri-gram ($n = 3$) features are I s, spe, eak, k E, etc.

The character n-grams can span across word boundaries and measure the sentence in different granularity by using different order n. Character n-grams of small n may capture phonotactics, morpheme (the smallest unit of a word) information, prefix and suffix usage (Tsur and Rappoport, 2007), and even some spelling errors, while those of large n may capture the whole word characteristics (Ahn, 2011). We experimented with character n-grams in various settings as in that for word n-grams.

POS n-gram The sequence of part-of-speech (POS) tags for adjacent words of length n.

The POS tags indicate the syntactic or morphological category of words. The Penn Treebank POS tag set (Mitchell Marcus, 2012) was used. For example, VBP stands for non-third-person singular present verb, PRP stands for personal pronoun, and NNP for singular proper noun. Therefore, the example POS bi-grams for the same sentence I speak English are PRP VBP and VBP NNP. We experimented POS n-grams of different orders ($n = 2$ to 5). Inspired by the study of Bykh and Meurers (2012), we adopted three different types of POS n-grams (Table 7.6).

The pure POS feature uses only POS tags without lexical information. The POS+Word feature is the lexicalized POS n-gram, binding the words to their POS tags. The hybrid feature replaces all open-class words, e.g., nouns, verbs, adjectives, etc., with their POS tags to eliminate content-dependent information. Each of the three types of POS n-grams was experimented with different orders ($n = 2$ to 5). For hybrid POS n-grams, we also replaced different word classes and compared the results. By doing that we could achieve a clear understanding about which word class contributed most to the classification performance.

Table 7.6 *Example of three POS n-gram subtypes*

Subtypes	Description	Example
Pure POS	POS tags only	[PRP, VBP, DT, NN]
POS + Word	Lexical words and their POS tags	[I_PRP, drink_VBP, the_DT, water_NN]
Hybrid	Replace open-class lexical words (i.e., nouns, verbs, adjectives) with their POS tags	[I VBP the NN]

Parse Tree Output:	
<pre>(ROOT (S (NP (PRP I)) (VP (VBP speak) (NP (NNP English))))</pre>	
Unlexicalized PR:	Lexicalized PR:
$S \rightarrow NP + VP$ $NP \rightarrow PRP$ $VP \rightarrow VBP + NP$ $NP \rightarrow NNP$	$S \rightarrow NP + VP$ $NP \rightarrow PRP_I$ $VP \rightarrow VBP_speak + NP$ $NP \rightarrow NNP_English$

Figure 7.1 Example of unlexicalized and lexicalized version of production rule features (PR) of sentence “I speak English”.

Production rule The rewriting rule that specifies a symbol substitution for generating a new symbol sequence.

A production rule is extracted from a sentence parse tree under the framework of Context Free Grammar (CFG). For example, from the sentence `I speak English`, we can extract the production rule $S \rightarrow NP + VP$, which stands for the structure of a noun phrase NP (`I`) and a verb phrase VP (`speak English`). In addition to standard production rules, we also tested lexicalized production rules, with the corresponding words attached to each symbol, e.g., $S \rightarrow NP_I + VP_went$. Figure 7.1 demonstrates the parse tree for a sample sentence and the two versions of production rule features respectively.

Dependency The functional relationship between a pair of words, where one word is the head and the other is the dependent.

In the example sentence, `dobj(speak, English)` is a dependency relation, which denotes that `English` is the direct object (`dobj`) of the verb `speak`. Such feature can capture the relations between noncontiguous word pairs. We adopted Stanford typed dependency scheme (SD), and obtained the dependencies by converting constituency parse trees with heuristic rules (De Marneffe, MacCartney, and Manning, 2006).

We experimented with both lexicalized and unlexicalized dependencies. Figure 7.2 demonstrates the dependency parse of sentence `I come from western Europe` and the two versions of dependency features.

7.3.2 Machine Learning Techniques

We tried two popular classifiers in text classification: Naive Bayesian (NB) and Support Vector Machine (SVM) for NLI. The result showed that the latter

Dependencies Output:	
nsubj(come-2, I-1) root(ROOT-0, come-2) prep(come-2, from-3) amod(Europe-5, western-4) pobj(from-3, Europe-5)	
Unlexicalized Depd:	Lexicalized Depd:
nsubj prep amod pobj	nsubj_come_I prep_come_from amod_Europe_western pobj_from_Europe

Figure 7.2 Example of unlexicalized and lexicalized version of dependency features (Depd) on sentence “I come from western Europe”.

significantly outperformed the former in terms of overall accuracy. In a preliminary seven-way classification experiment where the data set size was 40K and the dimension of feature vector was 14K, SVM achieved an accuracy of 61% whereas NB achieved an accuracy of only 49%. In view of this, in the rest of this chapter, we only report the result from SVM classifier.

Support vector machine (SVM), having yielded the best performance in many NLP studies, is one of the most popular machine learning techniques in the field. A SVM classifier finds a separating hyperplane in the high-dimensional vector spaces of a given data set. A good hyperplane should have the largest distance to any data point of any class. The largest distance is termed as margin.

Given a data set $S = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$, the hyperplane that maximizes the margin is obtained by finding the hyperplane defined by $\langle \mathbf{w}, b \rangle$ that

$$\begin{aligned} & \text{minimizes: } \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{subject to constrain: } y_i(\mathbf{w}x_i - b) \geq 1 \end{aligned}$$

In the case where samples are not linearly separable, a kernel trick is adopted to project the data points into higher dimensional space where the data points are linearly separable. The inner products between high-dimensional data points can then be computed efficiently.

Most prior studies on NLI used SVM for classification (Koppel, Schler, and Zigdon, 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009; Wong and Dras, 2011; Kochmar, 2011; Al-Rfou, 2012; Bykh and Meurers, 2012; Tetreault, Blanchard, and Cahill, 2013). It can suit the unique properties of text classification: the feature vector dimension is usually many orders of magnitude higher than other classification tasks; the features are highly relevant across data set; sometimes they are linearly inseparable. These properties make

SVM suitable for text classification, which has also been attested by Joachims (1998).

SVM is originally designed for binary classification. However, for NLI, a multiclass classifier is needed. A number of strategies have been proposed for combining multiple binary SVMs to build a multiclass classifier (Bishop, 2006). The most common approach is to construct a set of SVMs for each individual class, where each SVM is trained using data from one class as positive examples and the rest of the classes as negative examples. Predictions for new inputs are made by choosing the class corresponding to the greatest margin. This is known as the one-versus-the-rest approach, which usually gives good results (Vapnik, 1998), and is the strategy used by LIBLINEAR (Fan et al., 2008) – the machine learning library that we used in our experiments.

7.3.3 Experiment Setup

We used the Stanford parser (Klein and Manning, 2003) to extract syntactic features. Following Bykh and Meurers (2012), we used the LibLinear SVM classifier (Fan et al., 2008) for NLI. To avoid selection bias, we performed fourfold cross-validation and reported the average accuracy at levels 1–3, 4–7, 8–11, and 12–16, respectively.

7.4 Results

First, we investigated the performance of individual feature types. We then investigated the combined effect of the features. Furthermore, we extracted the most distinguishing features and qualitatively analyzed these features with respect to the L1 backgrounds of the L2 learners. In general, our results show that lexical features (word and character n-grams) achieve higher classification accuracy than grammatical features (POS n-grams, production rules and dependencies). Nevertheless, as the proficiency level grows, the performance of lexical features drops while the performance of grammatical ones grows. This may imply that as students progress in L2, they use more complicated grammatical structures that show influence from their native languages.

7.4.1 Individual Features

This section presents our findings of the best settings for individual feature types in NLI and compares the accuracy of NLI across different individual feature types.

Effect of different feature settings In this section, we explore the effect of different feature settings on classification accuracy. For word n-grams,

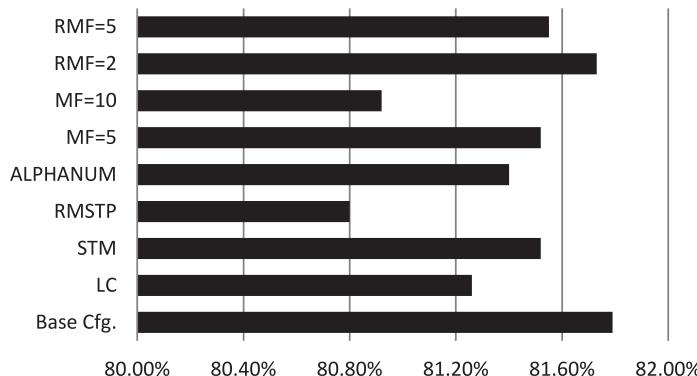


Figure 7.3 Performance of different configuration variations of word 1-gram feature at levels 1–3.

we first defined a baseline setting, which involved none of the normalization and filtering treatment mentioned in Section 7.3.1. We then applied one treatment at a time to observe its effect. Figure 7.3 presents the classification results of using each variation of word uni-gram features on the L2 proficiency group of levels 1–3 alone. The results in the other proficiency groups were similar.

We can see that the base configuration set achieved the highest accuracy among all variations. Any attempt to use less specific features degraded the performance. In particular, removing stop words resulted in the greatest drop of classification accuracy, which indicated that stop words were informative features for NLI. Nevertheless, setting the recurring minimum frequency to 2 ($RMF = 2$) attained nearly the same results as that of base configuration set, while reducing the feature dimensions by more than a half (14,921 versus 36,012). This meant the treatment was useful: it discarded features that applied to only one individual author and did not generalize to his or her L1 language group, thus reducing noisy features. As a result, in subsequent experiments that involved word n-gram features or any other lexicalized features, we set the RMF as 2.

Figure 7.4 shows the classification accuracies of different character 5-grams for the L2 proficiency groups of levels 1–3. Similar trends were observed on other n-grams and L2 proficiency groups. Again, normalization and filtering treatment resulted in a slight decrease in the classification accuracy across different L2 proficiency groups. This meant that for NLI on the data set of EFCAMDAT, more specific features led to higher classification accuracy.

Meanwhile, the optimal length for word/character n-gram features varied across different proficiency levels. In general, a larger n was preferred as the

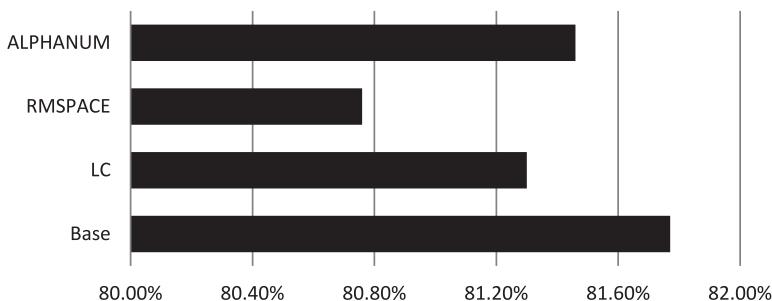


Figure 7.4 Performance of different variations of character 5-gram at levels 1–3.

proficiency level increased. This might be attributable to the fact that longer and more complex expressions were featured in the writings of advanced learners.

Figure 7.5 shows the NLI accuracies achieved by various POS features on the L2 proficiency group of levels 1–3. Again, the patterns on other L2 proficiency groups were similar. As we can see, the POS+Word subtype outperformed the other two subtypes. However, this advantage shrank as the order n grew. This was possibly because for lexicalized POS+Word features, uni-grams were already expressive and distinguishable, whereas higher-order n -grams suffered from data sparsity. On the other hand, for less specific POS features (pure POS and hybrid POS), higher-order POS sequences brought in more discriminatory information on native language backgrounds.

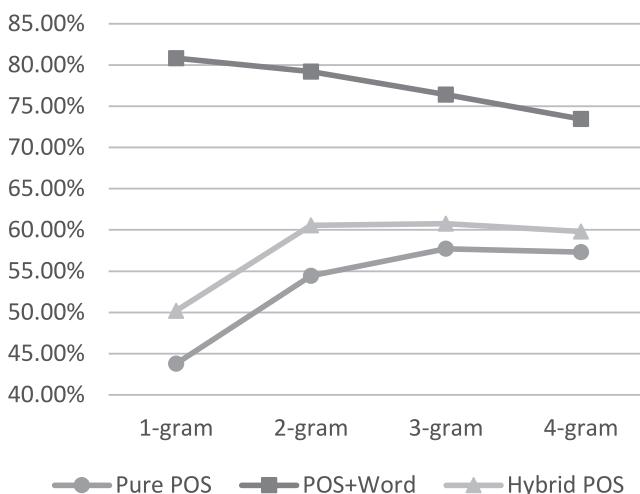


Figure 7.5 Performance of different POS n -grams at levels 1–3.

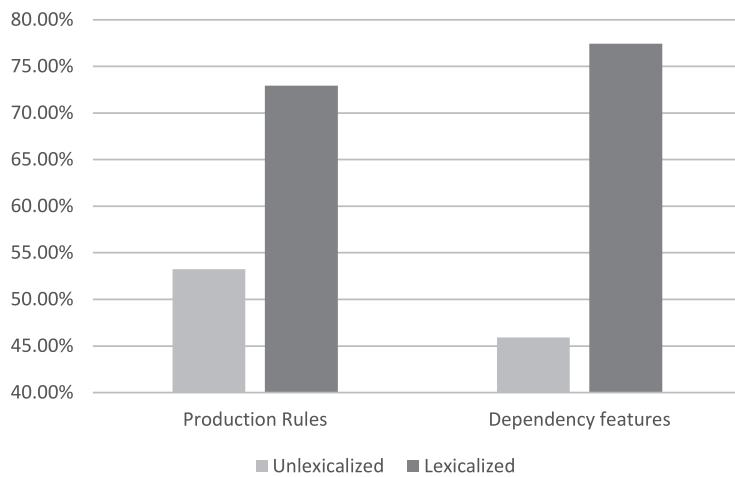


Figure 7.6 Performance of unlexicalized and lexicalized syntactic relational features at levels 1–3.

For production rules and dependencies, their lexicalized version also performed better than the unlexicalized one in most cases: Figure 7.6 demonstrates the classification accuracies of these two settings on the L2 proficiency group of levels 1–3. Similar results were found in other L2 proficiency groups.

Comparison across individual features This section compares the accuracy of NLI across individual features. Table 7.7 shows the accuracy of our NLI system when using each individual feature type alone. Here we report the results for the best configuration of the features only.

As shown in Table 7.7, lexical features (word and character n-grams) significantly outperformed syntactic ones (POS n-grams, production rules and dependencies) by up to 36%. Furthermore, the discriminatory effect of lexical features was more significant for beginners than for advanced learners. This might result from the fact that the former just started to learn morphology and

Table 7.7 *Performance of each individual feature type*

	1–3	4–7	8–11	12–16
Word	81.16%	79.17%	77.57%	63.12%
Char	81.19%	81.60%	79.30%	66.89%
POS	57.32%	62.37%	62.81%	56.29%
PR	51.88%	58.95%	60.32%	53.71%
Depd	45.16%	51.50%	55.06%	49.62%

lexicon, and were thus subject to more influence from L1. For syntactic features, the impact of L1 was clearer at medium than at low or high proficiency levels. This was probably because at the beginner levels, students were exposed to simple syntactic constructions that were relatively easy to learn, while by the time they got to the advanced levels, most students had a better grasp of grammar and their L1 background had less impact. POS tags were the most telling of the three syntactic features probably because they can tap into morpho-syntax as well as lexically driven patterns, whereas production rules and dependencies were too abstract.

7.4.2 Combination of Features

This section investigates the combined effect of the features in NLI. Table 7.8 shows the results of using all but one of the feature types. We can see that lexical features contributed to overall performance in almost all cases, with the only exception for character n-gram for levels 1–3. When combined with other feature types, syntactic features like production rules and dependencies played a less important role in NLI as the proficiency level of the students increased, whereas POS n-gram became more and more indispensable.

To conclude, we observe that most of the individual feature types had significant impact on the overall results. However, the impact of individual features changes across proficiency, indicating that the language of learners changes. It is also noticeable that the overall accuracy drops at the highest levels 12–16. This is probably due to the fact that, at most advanced levels, learner language is less influenced by L1 as learners acquire most of their target English L2. From a practical perspective, the results also suggest that when building NLI system for L2 data across different proficiency levels, we should consider different combination of features for different L2 proficiency levels.

Table 7.8 Accuracy gain (+%) or loss (−%) of leave-one-out experiments

Lvl	1–3	4–7	8–11	12–16
All	82.09%	82.54%	80.84%	69.50%
w/o Word	−0.66	−0.45	−0.37	−0.53
w/o Char	+0.68	−2.09	−2.11	−2.57
w/o POS	+0.73	+0.43	−0.42	−1.20
w/o PR	−0.18	−0.24	+0.48	+1.38
w/o Depd	+0.11	+0.24	+0.27	+0.40

Table 7.9 *The top 5 features for different feature types and proficiency levels.*
Please refer to the Penn Treebank POS tag set (Mitchell Marcus, 2012) and the Stanford parser (De Marneffe and Manning, 2008) for the meanings of POS tags, production rules and typed dependencies. Underscores “_” in character 6-grams represent a white space between words. COMMA and DOT refer to the punctuations of comma and dot.

Word n-grams (n = 1)	
Lvl 1–3	russia; brazil; china; moscow; paulo
Lvl 12–16	which; brazil; that; it; suitable
Char n-grams (n = 6)	
Lvl 1–3	Russia; m_Russ; China_; om_Bra; m_Bras
Lvl 12–16	which; brazil; becaus; As_for; _suita
POS n-grams (n = 2)	
Lvl 1–3	FW NNP; NNP FW; NNP NNP; MD VB; PRP MD
Lvl 12–16	COMMA PRP; COMMA IN; COMMA CC; NNS DOT; NN PRP
Production rules	
Lvl 1–3	$NP \rightarrow NNP + FW + NNP$; $VP \rightarrow MD + VP$; $VP \rightarrow MD + RB + VP$; $PP \rightarrow IN + NP$; $ADJP \rightarrow NP + JJ$
Lvl 12–16	$S \rightarrow PP + NP + VP$; $S \rightarrow CC + NP + VP$; $S \rightarrow S + CC + S$; $S \rightarrow SBAR + NP + VP$; $S \rightarrow ADVP + NP + VP$
Dependencies	
Lvl 1–3	neg; npadvmod; ccomp; det; prep_opposite
Lvl 12–16	prepc_as_for; prep_about; prep_of; prep_from; preconj

7.4.3 Qualitative Analysis

We selected up to 100 best-performing features for each feature type using the Information Gain (IG) (Yang and Pedersen, 1997) criteria. IG measures the information (in the number of bits) obtained for classification by knowing the presence or absence of a feature. Table 7.9 shows the top five features for word uni-grams, character 6-grams, POS bi-grams, production rules, and dependencies, respectively for the lowest proficiency group and the highest proficiency group. The top features with other configurations (e.g., n-grams of different orders) had similar trends and are not shown here.

As shown in Table 7.9, the most indicative features varied a lot from one proficiency level to another. Take word n-grams, for example: the best-performing features for the beginners were country names that expressed one’s L1 background explicitly. Proper names are replaced by function words such as *which* and *that* at higher levels, which indicates the growth of the learners’ grammatical knowledge. We also notice a shift from phrasal rules such as $NP \rightarrow NNP + FW + NNP$ at lower levels to sentence level rules such as

Table 7.10 *Example of unique punctuations used in particular class: Brazilians(br) make more mistakes of replacing a quote mark with an acute accent mark*

Lvl	Type	Feature	ru	cn	br	Example
1–3	Char.	't	4	0	238	“I don ’t”
8–16	Char.	's	3	0	338	“It 's”

$S \rightarrow PP + NP + VP$ at higher proficiency levels. These results suggested that features corresponding to more complex syntactic structures tended to be more informative for NLI on advanced learners.

We also performed a qualitative analysis of these representative features. Some of our findings are summarized in the following sections.

Punctuation and lexical preferences Results showed that Brazilians were more likely to mistake the acute accent mark ' (ASCII = 180) for the standard single quote ' (ASCII = 47). This phenomenon remained in the writings by Brazilians of higher English proficiency level, as is shown in the following table.

Chinese learners did not use the dash mark as frequently as Russian and Brazilian learners did, which is shown in Table 7.11. This might be attributed to the native language transfer: since the dash mark was rarely used in the written text in Chinese, Chinese learners used it less in their English writings.

Some phrases were used more frequently by the authors of a particular native language background, such as those listed in Table 7.12. These phrases were manually checked to ensure that they distributed across different topics and across different proficiency levels. They did occur more frequently in the writings by Russian or Chinese learners.

Clause-initial prepositional phrase (CIPP) A syntactic feature discriminating between Russians, Brazilians, and Chinese was the production rule PP-NP-VP, which involved sentences in which a prepositional phrase (PP) appears at the beginning of a clause as in (1).

Table 7.11 *Chinese Learners tend to underuse dash*

Lvl	Type	Feature	ru	cn	br
1–3	Word	—	2327	64	1535
8–16	Char.	—	1130	93	843

Table 7.12 *Example of phrases frequently used by particular class*

Lvl	Type	Feature	ru	cn	br	Example
8–16	Word	<i>as for me</i>	127	5	0	“As for me I prefer to eat at home”
8–16	Word	<i>to my mind</i>	80	1	1	“To my mind it is the most important thing for me.”
4–7	Word	<i>try my best</i>	1	105	0	“I will try my best”
4–7	Word	<i>so I</i>	595	1987	804	“So I decide to leave”
8–16	Word	<i>whats more</i>	6	89	0	“Whats more, there are too much oil of main course.”
8–16	Word	<i>have a try</i>	4	77	0	“I urge you to have a try”

- (1) 1 *in 18:00 o'clock* he has dinner
 2 *In the evening* he eats dinner at 18 o'clock
 3 *according to market research* our clients consider that our logo is old fashioned
 4 *opposite the window* there is a big TV

What was interesting about this feature was that the direction of correlation with national language changed across proficiency. At the early beginner levels 1–3, Chinese learners produced clause initial PPs much more often than Russians and Brazilians did: “ru”: 725, “cn”: 1672, “br”: 849. However, in late beginner/early intermediate levels 4–7, it was the Russians that were the most productive, followed by Brazilians: “ru”: 3626, “cn”: 1397, “br”: 2798. This trend seemed consolidated at intermediate and advanced levels 8–16: “ru”: 1584, “cn”: 247, “br”: 776.

It was reasonable to hypothesize that the shift in the use of this rule from early beginner levels was linked to qualitative changes in the way this rule was used. We therefore inspected the actual productions to gain some insight. We compared productions for early beginner levels, when the rule was used dominantly by Chinese learners, with productions from all the other levels when Russians and Brazilians used this rule more.

CIPP at Levels 1–3

At this level learners of all nationalities used clause-initial prepositional phrases to convey primarily temporal information, like the time of day and the day of week etc. as in (2).

- (2) 1 *In the afternoon* he goes shopping at 3 o'clock (**cn**)
 2 *On Sunday*, he goes to the park and meets friends, and *at half past eleven* he plays tennis with his friends (**cn**)
 3 *On Saturday at eleven-thirty* he was going to swim (**ru**)
 4 *In the evening at 6 o'clock* he eat dinner (**ru**)

But there were some differences in the phrases used by different learners. First of all, Russians tended to use more complex prepositional phrases with two points of time reference, e.g., day of the week and time as in (2)c-d. This pattern was not as productive with Chinese and Brazilian learners. Meanwhile, a fact that set Russian learners apart from Brazilians and Chinese was the production of a wider range of phrases that went beyond strict reference to time and location as indicated in the following examples.

- (3) 1 I think we have to say that *in success case* they will get additional bonuses (**ru**)
- 2 I want to learn English that *in future* I'll can understand other people (**ru**)
- 3 *With great pleasure* we inform our clients and shareholders of the change of company's logo (**ru**)
- 4 so *in the nearest time* we do not expect good growth (**ru**)
- 5 Workspaces are not clean and tidy : *in fact* they are messy (**ru**)

Second, Brazilians introduced many of their temporal phrases with *after* as in (4). In examples like (4) the temporal prepositional phrase established a sequential link between events (rather than pointing to a specific point in time).

- (4) 1 and *after breakfast* I go to work with my dad (**br**)
- 2 *after the movie* We have some coffee (**br**)

Meanwhile, Brazilians used prepositional phrases to mark not only time but also space/location as in (5). While we can find such uses in the productions of Chinese and Russian learners, Brazilians appeared much more likely to bring information on location at the beginning of their sentences.

- (5) 1 *On Park Road* there is a movie theater, *near the movie theater* there are many clothes stores and books stores (**br**)
- 2 and *in coffee room* there is one table (**br**)
- 3 In opposite you have one pharmacy, and *on the left* you have the a market (**br**)
- 4 There are two big windows and *opposite my bed* I put a table with a TV on it (**br**)

CIPP at Levels 4–16

As expected, learners used a wider range of prepositional phrases in higher proficiency levels, a fact reflecting their ability to better express themselves in their L2 English. The range of clause-initial prepositional phrases was extended for all three nationalities, as shown in (7).

- (6) 1 I can not go surfing because *for me* it is too scary (**ru**)
- 2 and *in the process* she changes size several times (**ru**)
- 3 but *at last* these girls decided not to attend (**cn**)

Reference to time and location remains dominant, but it involves more complex or abstract temporal meanings as indicated in (7).

- (7) 1 She set the conditions that *before the wedding* she will be traveling on the ship (**ru**)
- 2 I hope that *after all my efforts* I'll get a job at the company of my dream (**ru**)
- 3 so *until that time* it's sure of that I will share more time with my wife and my baby (**cn**)
- 4 but *at the end* everything was resolved (**br**)
- 5 I hope that *during my absence* you take care of everything (**br**)

The observations mentioned earlier indicate that our methodology can capture changing patterns in learner language across proficiency. To explain these changing patterns, we hypothesize that the overuse of the rule by Chinese learners at early beginner stages compensates for the absence of tense morphology in their grammar, an area that is known to be challenging for Chinese learners (Lardiere, 1998). A prediction of this hypothesis is that the drop in the use of PP preposing should correlate with increasing use and accuracy of verbal tense morphology. Regarding Russians, we hypothesize that preposing is due to transfer from L1 Russian where preposing is frequent for information structure (King, 1995). A prediction of this hypothesis is that the increase in PP preposing in Russian learners correlates with productive marking of information structure (e.g., higher accuracy in nominal anaphora, etc). Testing these predictions is beyond the scope of this chapter. However, the more general point is that NLI in a longitudinal corpus can capture both L1 effects as well as changing patterns across the learning trajectory, which can lead to new hypotheses for researchers in second language acquisition.

7.5 Conclusion

We developed a method for NLI that employs accurate machine learning (SVM) with a wide range of linguistic features (ranging from character features to syntactic dependencies) and applied this method to the newly developed, large EFCAMDAT corpus that, unlike previous learner corpora, provides longitudinal data at multiple proficiency levels. For the first time, we compared the performance of different feature types in NLI at different proficiency levels. We reported high overall accuracy of around 80% at low and medium proficiency levels and 70% at advanced levels. Our quantitative and a qualitative analysis of different features revealed that the top performing features differed from one proficiency level to another. Our linguistic analysis showed that our results can be of interest to research on L2 acquisition.

In the future, we plan to investigate NLI at finer-grained levels of proficiency and to integrate a wider range of nationalities, exploring strategies to deal with data sparsity. We also plan to develop new NLI methodology suitable for the analysis of large, longitudinal data, based on the insights gained in our experiments. Finally, we plan to conduct further linguistic evaluation of the data.

Acknowledgments

We thank EF Education First for providing the data, sponsoring the development of EFCAMDAT and the EF Research Lab for Applied Language Learning, the Isaac Newton Trust (Cambridge) for a grant supporting the development of EFCAMDAT, and finally the Royal Society, UK.

References

- Ahn, Charles S. (2011). “Automatically detecting authors’ native language”. Ph.D. thesis, Monterey, California. Naval Postgraduate School.
- Al-Rfou, Rami. (2012). “Detecting English Writing Styles For Non-native Speakers”. In: *arXiv preprint arXiv:1211.0498*.
- Bestgen, Yves, Sylviane Granger, Jennifer Thewissen, et al. (2012). “Error patterns and automatic L1 identification”. In: *Approaching language transfer through text classification*, pp. 127–153.
- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, pp. 338–339.
- Blanchard, Daniel, et al. (2013). “TOEFL11: A corpus of non-native English”. In: *Educational Testing Service*.
- Brooke, Julian, and Graeme Hirst (2011). “Native language detection with cheap learner corpora. In: *Conference of Learner Corpus Research (LCR2011)*.
- Bykh, Serhiy, and Detmar Meurers (2012). “Native Language Identification Using Recurring N-grams—Investigating Abstraction and Domain Dependence”. In: *Proceedings of COLING 2012: Technical Papers*, pp. 425–440.
- Charniak, Eugene, and Johnson, Mark. (2005). “Coarse-to-fine n-best parsing and Max-Ent discriminative reranking”. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 173–180.
- De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D Manning (2006). “Generating typed dependency parses from phrase structure parses”. In: *Proceedings of LREC*, Vol. 6, pp. 449–454.
- De Marneffe, Marie-Catherine, and Christopher D Manning (2008). “Stanford typed dependencies manual”. In: URL <http://nlp.stanford.edu/software/dependenciesmanual.pdf>.
- Estival, Dominique et al. (2007). “Author profiling for English emails”. In: *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING07)*, pp. 263–272.

- Fan, Rong-En et al. (2008). "LIBLINEAR: A library for large linear classification". In: *The Journal of Machine Learning Research* 9, pp. 1871–1874.
- Geertzen, Jeroen, Theodora Alexopoulou, and Anna Korhonen (2013). "Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT)". In: *in Proceedings of the 31st Second Language Research Forum (SLRF), Carnegie Mellon*. Cascadilla Proceedings Project.
- Graesser, Arthur C et al. (2004). "Coh-Metrix: Analysis of text on cohesion and language". In: *Behavior Research Methods, Instruments, & Computers* 36.2, pp. 193–202.
- Granger, Sylviane. (2003). "The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research". In: *Tesol Quarterly*, 37.3, pp. 538–546.
- Ionescu, Radu Tudor, Marius Popescu, and Aoife Cahill (2014). "Can characters reveal your native language? A language-independent approach to native language identification". In: *Proceedings of EMNLP, Octombrie*.
- Jarvis, Scott (2011). "Data mining with learner corpora". In: *A Taste for Corpora: In Honour of Sylviane Granger* 45.
- Jarvis, Scott, Yves Bestgen, and Steve Pepper (2013). "Maximizing Classification Accuracy in Native Language Identification". In: *NAACL/HLT 2013*, p. 111.
- Jarvis, Scott, and Scott A Crossley (2012). *Approaching Language Transfer through Text Classification: Explorations in the Detectionbased Approach*. Multilingual Matters.
- Joachims, Thorsten (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer.
- King, Tracy Halloway. (1995). *Configuring Topic and Focus in Russian*. CSLI Publications.
- Klein, Dan, and Christopher D Manning (2003). "Accurate unlexicalized parsing". In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 423–430.
- Kochmar, Ekaterina (2011). "Identification of a writer's native language by error analysis". Ph.D. thesis, Master's thesis, University of Cambridge.
- Koppel, Moshe, Jonathan Schler, and Kfir Zigdon (2005). "Automatically determining an anonymous authors native language". In: *Intelligence and Security Informatics*. Springer, pp. 209–217.
- Lardiere, Donna (1998). "Case and tense in the 'fossilized' steady state". In: *Second Language Research*, 14, pp. 1–26.
- Lavergne, Thomas, et al. (2013). "LIMSI Participation in the 2013 Shared Task on Native Language Identification". In: *NAACL/HLT 2013*, p. 260.
- Malmasi, Shervin and Mark Dras (2014a). "Arabic Native Language Identification". In: *ANLP 2014*, p. 180.
- (2014b). "Chinese Native Language Identification". In: *EACL 2014*, p. 95.
- Mitchell Marcus, Ann Taylor, Robert MacIntyre (2012). *Alphabetical list of part-of-speech tags used in the Penn Treebank Project*. URL: http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html (visited on 05/30/2013).
- Popescu, Marius and Radu Tudor Ionescu (2013). "The Story of the Characters, the DNA and the Native Language". In: *NAACL/HLT 2013*, p. 270.

- Swanson, Ben and Eugene Charniak (2012). “Native language detection with tree substitution grammars”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, pp. 193–197.
- Tetreault, Joel, Daniel Blanchard, and Aoife Cahill (2013). “A report on the first native language identification shared task”. In: *NAACL/HLT 2013*, p. 48.
- Tetreault, Joel R et al. (2012). “Native Tongues, Lost and Found: Resources and Empirical Evaluations in Native Language Identification”. In: *COLING*, pp. 2585–2602.
- Tomokyo, Laura Mayfield and Rosie Jones (2001). “You’re not from ‘round here, are you?: naive Bayes detection of non-native utterance text”. In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, pp. 1–8.
- Tsur, Oren and Ari Rappoport (2007). “Using classifier features for studying the effect of native language on the choice of written second language words”. In: *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Association for Computational Linguistics, pp. 9–16.
- Tsvetkov, Yulia et al. (2013). “Identifying the L1 of non-native writers: the CMU-Haifa system”. In: *NAACL/HLT 2013*, p. 279.
- Vapnik, V.N (1998). *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, pp. 437–438.
- Wong, Sze-Meng Jojo, and Mark Dras (2009). “Contrastive analysis and native language identification”. In: *Proceedings of the Australasian Language Technology Association Workshop*. Citeseer, pp. 53–61.
- (2011). “Exploiting parse structures for native language identification”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1600–1610.
- Wong, Sze-Meng Jojo, Mark Dras, and Mark Johnson (2011). “Topic modeling for native language identification”. In: *Proceedings of the Australasian Language Technology Association Workshop*, pp. 115–124.
- (2012). “Exploring adaptor grammars for native language identification”. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pp. 699–709.
- Yang, Yiming and Jan O Pedersen (1997). “A comparative study on feature selection in text categorization”. In: *International Conference on Machine Learning*. Morgan Kaufmann Publishers, Inc., pp. 412–420.

8 Evaluating Language Acquisition Models: A Utility-Based Look at Bayesian Segmentation

Lisa Pearl and Lawrence Phillips

Abstract

Computational models of language acquisition often face evaluation issues associated with unsupervised machine learning approaches. These acquisition models are typically meant to capture how children solve language acquisition tasks without relying on explicit feedback, making them similar to other unsupervised learning models. Evaluation issues include uncertainty about the exact form of the target linguistic knowledge, which is exacerbated by a lack of empirical evidence about children’s knowledge at different stages of development. Put simply, a model’s output may be good enough even if it does not match adult knowledge because children’s output at various stages of development *also* may not match adult knowledge. However, it is not easy to determine what counts as “good enough” model output. We consider this problem using the case study of speech segmentation modeling, where the acquisition task is to segment a fluent stream of speech into useful units like words. We focus on a particular Bayesian segmentation strategy previously shown to perform well on English, and discuss several options for assessing whether a segmentation model’s output is good enough, including cross-linguistic utility, the presence of reasonable errors, and downstream evaluation. Our findings highlight the utility of considering multiple metrics for segmentation success, which is likely also true for language acquisition modeling more generally.

8.1 Introduction

A core issue in machine learning is how to evaluate unsupervised learning approaches (von Luxburg, Williamson, & Guyon, 2011), since there is no *a priori* correct answer the way that there is for supervised learning approaches. Computational models of language acquisition commonly face this problem because they attempt to capture how children solve language acquisition

tasks without explicit feedback, and so typically use unsupervised learning approaches. Moreover, evaluation is made more difficult by uncertainty about the exact nature of the target linguistic knowledge and a lack of empirical evidence about children's knowledge at specific stages in development. Given this, how do we know that a model's output is "good enough"? How should success be measured? To create informative cognitive models of acquisition that offer insight into how children acquire language, we should consider how to evaluate acquisition models appropriately (Pearl, 2014; Phillips, 2015; Phillips & Pearl, 2015b).

As a case study, we investigate the initial stages of speech segmentation in infants, where a fluent stream of speech is divided into useful units, such as words. For example, the acoustic signal transcribed via IPA as /ajlʌvðizpəŋgwinz/ (*I love these penguins*) might be segmented as /aj lʌv ðiz pəŋgwinz/ (*I love these penguins*). A particular Bayesian segmentation strategy has been shown to be quite successful on English (Goldwater, Griffiths, & Johnson, 2009; Pearl, Goldwater, & Steyvers, 2011; Phillips & Pearl, 2012, 2014a, 2014b; Phillips, 2015; Phillips & Pearl, 2015b), particularly when cognitive plausibility considerations have been incorporated at both the computational and algorithmic levels of Marr (1982). One way to evaluate if this strategy is "good enough" is to see how it fares cross-linguistically. This is based on the premise that core aspects of the language acquisition process – such as the early stages of segmentation occurring in six- to seven-month-olds (Thiessen & Saffran, 2003; Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) – are universal. So, a viable learning strategy for early segmentation should succeed on any language infants encounter.

Traditionally, a segmentation model's output has been compared against a "gold standard" derived from adult orthographic segmentation (e.g., Brent, 1999; M. Johnson, 2008; Goldwater et al., 2009; Blanchard, Heinz, & Golinkoff, 2010; M. Johnson, Demuth, Jones, & Black, 2010; M. Johnson & Demuth, 2010; Pearl et al., 2011; Lignos, 2012; Fourtassi, Börschinger, Johnson, & Dupoux, 2013). Notably, orthographic segmentation assumes the desired units are orthographic words. However, if we look at the world's languages, it becomes clear that it is also useful to identify morphemes – the smallest meaningful linguistic units – particularly for languages with regular morphology. That is, infants might reasonably segment morphemes from fluent speech rather than entire words. Notably, models that identify subword morphology are penalized by the gold standard evaluation, and this highlights the need for a more flexible metric of segmentation performance. More generally, a segmentation strategy that identifies units useful for later linguistic analysis should not be penalized.

Still, how do we know that the segmented units are truly useful? If we believe that the output of early segmentation scaffolds later acquisition

processes, a useful segmentation output should simply enable these later processes to successfully occur (Phillips & Pearl, 2015a). For example, one goal of early segmentation is to generate a proto-lexicon in order to bootstrap language-specific segmentation cues like stress pattern (e.g., in English, words in child-directed speech tend to begin with stress [Swingley, 2005], and the same is true for child-directed German and Hungarian [Phillips & Pearl, 2015a]). Does the inferred proto-lexicon of units yield the appropriate language-specific cue? As another example, infants begin to learn mappings from word forms to familiar objects as early as six months (Tincoff & Jusczyk, 1999; Bergelson & Swingley, 2012; Tincoff & Jusczyk, 2012). Can the inferred proto-lexicon be used successfully for this process?

In the remainder of this chapter, we first review relevant aspects of infant speech segmentation, including what is known about the developmental trajectory, the cues infants are sensitive to, and how infants perceive the input. This forms the empirical basis for the modeled Bayesian segmentation strategy, which we then discuss in terms of its underlying generative assumptions and the algorithms used to carry out its inference. We turn then to the cross-linguistic evaluation of this strategy over input derived from child-directed speech corpora in seven languages from the CHILDES database (MacWhinney, 2000): English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. This section includes comparison to the gold standard as well as a more flexible metric derived from the gold standard that's consistent with children's imperfect early segmentation behavior. We find that the Bayesian strategy seems to be "good enough" cross-linguistically. This is especially true once we use this more nuanced output evaluation that considers potentially useful nonword units valid. This serves as a general methodological contribution about the definition of segmentation success, especially when we consider that useful units may vary across the world's languages.

We conclude with an evaluation metric that is even more utility-driven: whether the output of the segmentation strategy is helpful for subsequent acquisition processes, such as inferring a language-specific stress-based segmentation cue and learning early word-meaning mappings. Interestingly, just because a strategy yields more accurate segmentations when compared against the gold standard does not mean it is always more useful for subsequent acquisition processes. This underscores the value of considering multiple metrics for segmentation success, in addition to the traditional comparison against the gold standard.

8.2 Early Speech Segmentation

Segmentation is not easy – words blur against one another, making speech more like a stream of sound rather than something divided into discrete,

separable chunks (Cole & Jakimik, 1980). Yet, speech segmentation is one of the first tasks infants accomplish in their native language, and the resulting segmented units underlie subsequent processes such as learning word meanings, syntactic categories, and syntactic structure. In order to accurately model the segmentation process, we need to know the empirical data that form the basis for decisions regarding the model's learning assumptions, input, inference, and evaluation.

8.2.1 When Does Early Speech Segmentation Begin?

The first behavioral evidence for speech segmentation in infants comes at six months (Bortfeld et al., 2005), when infants seem to know a small set of very frequent words (Bergelson & Swingley, 2012). They recognize these words in speech and can use them to segment new utterances. Between seven and nine months, infants learn to utilize language-specific cues such as stress pattern (Jusczyk, Cutler, & Redanz, 1993; Jusczyk, Houston, & Newsome, 1999; Thiessen & Saffran, 2003), phonotactics (Mattys, Jusczyk, & Luce, 1999), allophonic variation (Hohne & Jusczyk, 1994; Jusczyk, Hohne, & Baumann, 1999), and coarticulation (E. Johnson & Jusczyk, 2001). These language-specific cues are typically more reliable than language-independent cues such as transitional probability between syllables. However, it turns out that infants around seven months old prefer to rely on transitional probability information alone rather than language-specific cues such as stress patterns (Thiessen & Saffran, 2003), even though transitional probability is a less reliable cue. Neuroimaging evidence from neonates suggests that this sensitivity to statistical cues like transitional probabilities is present at birth (Teinonen, Fellman, Näätänen, Alku, & Huotilainen, 2009). This in turn suggests that the initial stages of speech segmentation rely on cues that are independent of the particular language being segmented (e.g., the process of tracking transitional probabilities does not vary from language to language, though the probabilities themselves clearly do). It is only after this initial stage that infants harness the more powerful language-dependent cues that do vary between languages (e.g., the specific stress-based segmentation cues that differ between English and French). So, a model of early speech segmentation should also likely rely only on language-independent cues.

8.2.2 The Unit of Infant Speech Perception

The basic unit of infant speech perception has been a source of significant debate for some time (see Phillips, 2015 and Jusczyk, 1997 for a more detailed review of this debate). Experimental studies have typically focused on whether the basic representational unit for infants is syllabic or segmental (Jusczyk

& Derrah, 1987; Bertonicini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Bijeljac-Babic, Bertonicini, & Mehler, 1993; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995; Eimas, 1999). Jusczyk (1997, p. 115) summarizes several studies by noting that “there is no indication that infants under six months of age represent utterances as strings of phonetic segments.” Instead, evidence for segmental representations of speech is mostly present in infants older than six months: infants first begin to ignore vowel contrasts that aren’t relevant for their native language around six months (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994), while irrelevant consonant contrasts are ignored between eight and twelve months (Werker & Tees, 1984; Werker & Lalonde, 1988; Best, McRoberts, & Sithole, 1988; Best, McRoberts, LaFleur, & Silver-Isenstadt, 1995).

More generally, as infants get older, they are better able to represent information about segments; in contrast, they appear relatively comfortable with syllables from early on. This can be seen in infants’ ability to track statistical relationships: while transitional probabilities over syllables can be tracked at birth (Teinonen et al., 2009), transitional probabilities over segments first seem to occur around nine months (Mattys et al., 1999). Given this, a reasonable assumption for a model meant to capture segmentation strategies being used by six-month-olds would be that the input is perceived as a stream of syllables.

8.2.3 *Constraints on Infant Inference*

One thing that makes child language acquisition so impressive is that it is accomplished despite the many cognitive constraints imposed by the developing brain. Though there has been increasing interest in cognitively constrained language acquisition models (e.g., Anderson, 1990; Shi, Griffiths, Feldman, & Sanborn, 2010; Bonawitz, Denison, Chen, Gopnik, & Griffiths, 2011; Pearl et al., 2011; Phillips & Pearl, 2015b), there isn’t very much experimental evidence to suggest exactly what kinds of constraints should be imposed. There are many possibilities, but we focus on three that have been incorporated into past acquisition models and that seem reasonable starting points: online processing, nonoptimal decision-making, and recency effects.

Online processing refers to the idea that data are processed as they are encountered, rather than being stored in explicit detail for later batch processing. It is generally accepted that this is a reasonable constraint for human language processing, and commonly used as justification that a model operates at the algorithmic level in the sense of Marr (1982), rather than being idealized (e.g., Lignos & Yang, 2010; Pearl et al., 2011; Lignos, 2012; Phillips & Pearl, 2014b, 2014a, 2015b). So, this is likely reasonable to incorporate into infant inference.

For decision-making, experimental evidence from infants and children suggest that they do not always choose the highest probability option available, which would be considered the optimal decision (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009; Davis, Newport, & Aslin, 2011; Denison, Bonawitz, Gopnik, & Griffiths, 2013). Instead, children sometimes appear to probabilistically sample the available options (Davis et al., 2011; Denison et al., 2013). Other times, they appear to simply generalize to a single option, which might in fact be a lower probability option (Köpcke, 1998; C. H. Kam & Newport, 2005; C. L. H. Kam & Newport, 2009). This suggests that infant inference may involve nonoptimal decision-making.

With respect to memory constraints, experimental evidence suggests that a recency bias exists in infants (Cornell & Bergstrom, 1983; Gulya, Rovee-Collier, Galluccio, & Wilk, 1998; Rose, Feldman, & Jankowski, 2001), where the most recently encountered data have privileged status. So, this is also reasonable to incorporate into infant inference.

8.2.4 What We Know about Segmentation Output

As mentioned before, one empirical checkpoint for segmentation is that a strategy ought to be successful for any human language. Beyond that, we also have some evidence about the kinds of errors children produce – this underscores that successful early segmentation does not necessarily mean adultlike segmentation. For example, Brown (1973) and Peters (1983) find that even three-year-old children still produce missegmentations. These errors can be broadly split into two types: function word undersegmentations (e.g., *that'sa, it'sa*) and function word oversegmentations (e.g., *a nother, be have*). So, segmentation error patterns may provide a useful qualitative benchmark for model output, and have been previously used this way (Lignos, 2012; Phillips & Pearl, 2012, 2015b).

8.3 A Bayesian Segmentation Strategy

Bayesian segmentation strategies combine the prior probability of a potential segmentation s for an utterance u with the likelihood in order to generate the posterior probability of s ($P(s|u)$) using Bayes' rule, as shown in (8.1).

$$P(s|u) \propto P(s)P(u|s) \tag{8.1}$$

The Bayesian strategy we investigate builds off of two fundamental insights about the infant's inferred proto-lexicon, both of which were used in an earlier segmentation strategy by Brent (1999). First, frequent words should be preferred over infrequent words. Second, shorter words should be preferred over longer words. These parsimony biases were incorporated by Goldwater et al.

(2009) into a Bayesian strategy that used a Dirichlet Process (Ferguson, 1973) to determine the prior probability of a segmentation.

The Dirichlet Process (DP) is a nonparametric stochastic process resulting in a probability distribution often used in Bayesian modeling as a prior because it has properties well suited to language modeling. First, because it is nonparametric, it does not need to prespecify the number of items (e.g., word types) that might be encountered. Second, the DP facilitates “rich-get-richer” behavior, where frequent items are more likely to be encountered later. This is useful because word frequencies in natural languages tend to follow a power-law distribution that the DP naturally reproduces due to this behavior (Goldwater, Griffiths, & Johnson, 2011).

Goldwater et al. (2009) implemented two versions of the DP segmentation strategy that differed in their generative assumptions. Both versions use the likelihood function, i.e., the probability of an utterance given its segmentation $P(u|s)$, to simply rule out potential segmentations that do not match the observed utterance. For example, a possible segmentation /ðə pɛŋgwɪn/ (*the penguin*) does not match an observed speech stream /ðəkɪrɪ/ (*the kitty*) when the possible segmentation is concatenated (*thepenguin*=*thekitty*). So, this possible segmentation would have a likelihood of 0. In contrast, the possible segmentation /ðəki ri/ (*theki tty*) does match (*thekitty*=*thekitty*), and so would have a likelihood of 1. Where the DP strategy versions differ is how the prior probability for a segmentation is determined. We describe each version in turn before reviewing the inference algorithms paired with each generative model.

8.3.1 DP Segmentation: Unigram Assumption

The first version of the DP segmentation strategy uses a unigram language model (DP-Uni), with the modeled learner naively assuming that each word is chosen independently of the words around it. The prior of the potential segmentation is calculated using this generative assumption. To do this, the model must define the probability of every word in the utterance, and so a DP-Uni learner assumes that for any utterance, each segmented word w_i is generated by the following process:

1. If w_i is not a novel lexical item, choose an existing form ℓ for w_i .
2. If w_i is a novel lexical item, generate a form (e.g., the syllables $x_1 \dots x_M$) for w_i .

Because the model does not know whether w_i is novel, it has to consider both options when calculating the probability of the word. We note that deciding whether w_i is novel is not the same as deciding whether the form of w_i has been previously encountered. As an example, consider the first time the modeled learner encounters the sequence /et/ (such as in the word *ate*). Because

$count_{/et/} = 0$, the word must be novel ($count_{/et/}$ then = 1). Now, suppose the same sequence is encountered again, but from the word *eight*. The learner must decide if this sound sequence is a second instance of the /et/ it saw before (*ate*) or instead the first instance of a novel /et/ type (such as in the word *eight*). In the first case, the count might be updated to $count_{/et/} = 2$; in the second case, the counts might be updated to $count_{/et/} = 1$ and $count_{/et/2} = 1$. This particular aspect of the DP distribution naturally allows for the existence of homophones (e.g., *ate/eight*) without requiring any additional machinery.

Returning to the DP generation process, generating either non-novel or novel items is fundamental to the DP. In a classic DP, the probability of generating a non-novel item is proportional to the number of times that item has been previously encountered. This is shown in (8.2), where n_ℓ refers to the number of times lexicon item ℓ has been seen in the set of words previously encountered, denoted as w_{-i} . In the denominator, i represents the total number of words encountered thus far, including the word previously under consideration. Because the current word is not included in n_ℓ , 1 is subtracted from it.

$$P(w_i = \ell, w_i \neq \text{novel} | w_{-i}) = \frac{n_\ell}{i - 1 + \alpha} \quad (8.2)$$

Equation (8.2) gives higher probability to word types that have been encountered before. So, the more a word type is encountered by the modeled learner, the more often the modeled learner will prefer it in the future. This will end up generating the power-law frequency distribution found in natural languages.

When the DP instead generates a novel word, the word is not represented as an atomic whole, but rather constructed from its individual parts, such as syllables. To model this, we make use of the P_0 in (8.3) to describe the probability that any word might be made up of a particular string of subword units $x_1 \dots x_M$. Each subword unit x_j is generated in turn for all M units in the word, with the probability of x_j treated as a uniform choice over all possible subword units in the corpus.¹

$$P_0(w_i = x_1 \dots x_M) \propto \prod_{j=1}^M P(x_j) \quad (8.3)$$

The probability of generating a novel item in a DP is weighted by the free model parameter α , also known as the DP concentration parameter. This parameter has an intuitive interpretation, where higher values of α lead to a preference

¹ The model additionally includes the generation of word and utterance boundaries, with a word ending with some probability $p_{\#}$ and an utterance ending with some probability $p_{\$}$. See Goldwater et al. (2009) for discussion of these model components in the unigram and bigram versions of this strategy.

for generating novel words in the proto-lexicon.² The full probability of generating a novel word is therefore described by equation (8.4).

$$P(w_i = \ell, w_i = \text{novel} | w_{-i}) = \frac{\alpha P_0(w_i = x_1 \dots x_M)}{i - 1 + \alpha} \quad (8.4)$$

Both equations 8.2 and 8.4 can be combined to generate the full probability of a word being produced either as a non-novel or novel item.

$$P(w_i = \ell | w_{-i}) = \frac{n_\ell + \alpha P_0(w_i = x_1 \dots x_M)}{i - 1 + \alpha} \quad (8.5)$$

8.3.2 DP Segmentation: Bigram Assumption

The second version of the DP segmentation strategy uses a bigram language model (DP-Bi), with the learner assuming (slightly less naively) that each word is chosen based on the word preceding it. Goldwater et al. (2009) model this using a hierarchical Dirichlet Process (Teh, Jordan, Beal, & Blei, 2006), with the generative process selecting bigrams, words, and subword units as follows:

1. If the pair $\langle w_{i-1}, w_i \rangle$ is not a novel bigram, choose an existing form ℓ for w_i from those that have been previously generated after w_{i-1} .
2. If the pair $\langle w_{i-1}, w_i \rangle$ is a novel bigram:
 - If w_i is not a novel lexical item, choose an existing form ℓ for w_i .
 - If w_i is a novel lexical item, generate a form $(x_1 \dots x_M)$ for w_i .

As with the DP-Uni model, the DP-Bi model must make a decision between an item being novel or not; the main difference is that the DP-Bi model considers bigrams first. If a bigram is not novel, the DP-Bi model gives higher probability to bigrams that have been encountered before. If a bigram is instead novel, then the individual lexical item (the second word in the bigram) must also be generated. This is done with a DP in the same fashion as the unigram DP, making this a hierarchical DP. The probability of any bigram $\langle w_{i-1}, w_i \rangle$ is then determined using equations (8.6), (8.7), and (8.3).

$$P(\langle w_{i-1}, w_i = \ell \rangle | w_{-i}) = \frac{n_{\langle w_{i-1}, w_i = \ell \rangle} + \beta P_1(w_i = \ell | w_{-i})}{n_{w_{i-1}} + \beta} \quad (8.6)$$

Equation (8.6) calculates the probability of the bigram $\langle w_{i-1}, w_i \rangle$, given that the second word of the bigram w_i takes the form ℓ and considering all the words observed previously except w_i , denoted by w_{-i} . This includes the number of times ℓ appears as the second word of bigrams beginning with word w_{i-1} ($n_{\langle w_{i-1}, w_i = \ell \rangle}$), as well as the total number of bigrams beginning with w_{i-1} ,

² A more thorough treatment of the role of various model parameters for both the unigram and bigram DP segmentation models can be found in Goldwater et al. (2009).

denoted by $n_{w_{i-1}}$. The concentration parameter β determines how often a novel second word is expected, with higher values indicating a general preference for more novel bigrams.

Equation (8.7) describes the process for generating a novel second word in the bigram.

$$P_1(w_i = \ell | w_{-i}) = \frac{t_{w_i=\ell} + \gamma P_0(w_i = x_1 \dots x_M)}{t + \gamma} \quad (8.7)$$

A novel second word with form ℓ is based on the number of times any bigram with second word ℓ has been generated, $t_{w_i=\ell}$. The total number of novel bigrams is represented by t . The concentration parameter γ determines how often this novel second word is itself a novel lexical item, constructed from its constituent subword units $x_1 \dots x_M$ using P_0 , as in Equation (8.3).

8.3.3 DP Segmentation Inference

Pearl et al. (2011) used a variety of inference algorithms with these two DP segmentation strategies, including both idealized and constrained procedures. Idealized inference procedures provide a computational-level analysis in the sense of Marr (1982), and offer a best-case scenario of how useful the learning assumptions of the model are. Constrained inference procedures provide a more algorithmic-level analysis in the sense of Marr (1982). They in turn offer a more cognitively plausible assessment of how useable the learning assumptions are by humans, who have cognitive limitations on their inference capabilities (particularly infants). Here we focus on one inference algorithm of each kind.

8.3.3.1 Idealized Inference The original inference algorithm used by Goldwater et al. (2009) for DP segmentation was Gibbs sampling (Geman & Geman, 1984), a batch procedure commonly used for idealized inference, due to its guaranteed convergence on the optimal solution given the model constraints. Gibbs sampling initializes the model parameters (in this case potential word boundaries) and then updates each parameter value one at a time, conditioned on the current value of all other parameters. This process is repeated for a number of iterations until convergence is reached (e.g., Goldwater et al., 2009 and Pearl et al., 2011 used 20,000 iterations).

For DP segmentation, each possible boundary location is a parameter that either has a boundary or does not. Boundaries are initialized randomly, and the inference procedure goes through each boundary location in the corpus, deciding whether to place/remove a boundary, given the other current boundary locations. In particular, for each possible boundary, there is a choice between creating a single word (H_0) or two words (H_1) out of the two adjoining pieces. For example, H_0 might be /ðəkɪrɪ/ (*the kitty*) while H_1 is /ðə kɪrɪ/ (*the kitty*)

for the potential boundary location between /ðə/ and /kɪrɪ/. The probability of inserting a boundary (H_1) can be defined as the normalized probability of H_1 , shown in (8.8), with $P(H_0)$ and $P(H_1)$ defined by the DP-Uni or DP-Bi generative models:

$$\text{Normalized } P(H_1) = \frac{P(H_1)}{P(H_1) + P(H_0)} \quad (8.8)$$

The inference procedure then probabilistically selects either H_0 or H_1 , based on their normalized probabilities. If no boundary is placed (H_0), only a single word has to be generated; if a boundary is placed (H_1), two words have to be generated. Intuitively, the model may prefer either H_0 because it requires fewer words or H_1 because it requires shorter words. The exact trade-off depends on the model parameters and, most importantly, on the frequency each word (or bigram) is currently perceived to have.

8.3.3.2 Constrained Inference Pearl et al. (2011) described several inference procedures that incorporate one or more of the cognitive limitations relevant for infant speech segmentation mentioned before: (1) online processing, (2) nonoptimal decision-making, and (3) a recency bias. We focus on the one that incorporates all three of these constraints to some degree (called the DCMC constrained learner by Pearl et al., 2011). This inference procedure performs inference online, segmenting each utterance as it is encountered. It can also be thought to involve nonoptimal decision-making because it probabilistically samples whether to insert a boundary rather than always selecting the highest probability option.³

It additionally uses the Decayed Markov Chain Monte Carlo method (Marthi, Pasula, Russell, & Peres, 2002) to implement a recency bias. In particular, similar to the idealized inference procedure, it samples individual boundary locations and updates them conditioned on the value of all other currently encountered potential boundaries. However, instead of sampling all currently known potential boundaries equally, the locations to sample are selected based on a decaying function anchored from the most recently encountered potential boundary location (at the end of the current utterance). The probability of sampling potential boundary b_a , which is a potential boundaries away from the end of the current utterance, is determined by (8.9):

$$P(b_a) = \frac{a^{-d}}{\sum a_i^{-d}} \quad (8.9)$$

³ Unlike Gibbs sampling, which also probabilistically samples whether to insert a boundary, there is no guarantee of convergence on the optimal solution.

Table 8.1 *Summary of the syllabified child-directed speech corpora, including the CHILDES database corpora they are drawn from (Corpora), the age ranges of the children they are directed at (Age range), the number of utterances (# Utt), the number of unique syllables (# Syl types), the average number of syllables per utterance (Syls/Utt), and the probability of a word boundary appearing between syllables (B Prob).*

Language	Corpora	Age range	# Utt	# Syl types	Syls/Utt	B Prob
English	Brent	0;6–0;9	28391	2330	4.16	76.26
German	Caroline	0;10–4;3	9378	1682	5.30	68.60
Spanish	JacksonThal	0;10–1;8	16924	522	4.80	53.93
Italian	Gervain	1;0–3;4	10473	1158	8.78	49.94
Farsi	Family, Samadi	1;8–5;2	31657	2008	6.98	43.80
Hungarian	Gervain	1;11–2;11	15208	3029	6.30	51.19
Japanese	Noji, Miyata, Ishii	0;2–1;8	12246	526	4.20	44.12

The parameter d implements the recency effect: larger values of d indicate stronger biases, concentrating the boundary sampling efforts on more recently encountered data. We discuss results using $d = 1.5$, which implements a strong recency bias: using this d value, Phillips and Pearl (2015b) found that 83.6% of sampled boundaries occurred in the current utterance in their corpus of English child-directed speech, 11.8% in the previous utterance, and only 4.6% in any other previous utterance.

8.4 How Well Does This Work Cross-Linguistically?

8.4.1 Cross-Linguistic Corpora

Phillips and Pearl (2014a, 2014b) evaluated the DP segmentation strategy on seven languages: English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. These languages vary in many ways, including their morphology: some are more agglutinative and have more regular morphology systems (Hungarian, Japanese) while the others are fusional to different degrees and have less regular morphological systems (English, German, Spanish, Italian, and Farsi). Syllabified child-directed speech corpora were derived from the CHILDES database (MacWhinney, 2000; Gervain & Erra, 2012; Phillips & Pearl, 2015b),⁴ and relevant summary statistics for them are shown in Table 8.1.

We can make a few observations. First, not all languages had corpora available of speech directed at children younger than a year old, so the age range does vary. Second, the corpora vary in size, though this does not appear to

⁴ See Phillips (2015) for details of this process.

negatively impact the results for smaller corpora. Third, the number of unique syllables in each corpus varies considerably by language (e.g., Spanish: 522, Hungarian: 3,029). While some of this variation is due to the size of the corpus itself (more utterances allow more syllable types to appear), there are also language-specific phonotactic restrictions on syllables that impact the number of syllable types observed. For example, in Japanese only the phoneme /N/ may appear after a vowel, and in Spanish only the phoneme /s/ can appear as the second consonant in a coda. In contrast, languages such as English, German, and Hungarian allow much more complex syllable types (e.g., consider the English coda in *warmth*, /ɪmθ/).

The average number of syllables per utterance also varies, which is partially due to speech directed at older children containing longer utterances (e.g., Farsi utterances have 6.98 syllables per utterance and are directed at children up to age five). However, the number of syllables per utterance is also impacted by the number of syllables per word – languages that tend to be more monosyllabic will tend to have fewer syllables per word, and so their utterances will tend to have fewer syllables as well. Boundary probability captures this monosyllabic tendency, where higher probabilities indicate that syllables tend to be followed by boundaries (i.e., words are more likely to be monosyllabic). For example, the English and German data have higher boundary probabilities than the other languages, and therefore tend to have more monosyllabic words.

8.4.2 Gold Standard Evaluation

8.4.2.1 Evaluation Metrics We first present the DP segmentation’s ability to match the gold standard, the adult-level knowledge typically represented via the orthographic segmentation. There are multiple units a segmentation can be measured on: word tokens, lexical items, and word boundaries. For example, the utterance *The penguin is next to the kitty* might be segmented as *The penguin is next to the kitty*. There are seven word tokens (individual words) in the original sentence, but the segmentation only identifies three of those tokens (*is*, *the*, and *kitty*). The same utterance has six lexical items, which correspond to the unique words: *the*, *penguin*, *is*, *next*, *to*, *kitty* (*the* appears twice). Again, the segmentation only correctly identifies three lexical items (*is*, *the*, and *kitty*). This utterance also has six word boundaries (excluding the utterance boundaries) and the segmentation correctly identifies four of those (*thepenguin*, *is*, *is nextto*, *nextto the*, and *the kitty*). This highlights the differences between the units.

Word tokens are impacted by word frequency, while lexical items factor out word frequency. Measuring boundary identification tends to yield better performance than measuring word tokens or lexical items. Intuitively, this is because identifying a boundary requires the model to be correct only once. On contrast,

correctly segmenting words requires being correct twice, both in inserting the word-initial and word-final boundaries (unless the word is at an utterance edge).

No matter which unit we use for comparison, they are measured with the same metrics: precision and recall, which are often combined into a single summary statistic, the F-score. Precision captures how accurate the segmentation is: for each unit identified, is that unit correct in the segmentation? Recall captures how complete the segmentation is: for each unit that should have been identified, is that unit identified in the segmentation? In signal detection theoretic terms, these correspond to (8.10), which involve *Hits* (units correctly identified in the segmentation), *False Alarms* (units identified in the segmentation that aren't correct), and *Misses* (units not identified in the segmentation that are nonetheless correct).

$$\text{Precision} = \frac{\text{Hits}}{\text{Hits} + \text{False Alarms}} \quad \text{Recall} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}} \quad (8.10)$$

High precision indicates that when a unit is identified in a segmentation, it is often correct. High recall indicates that when a unit should be identified in a segmentation, it often is. Because both of these are desirable properties, they are often combined into the F-score via the harmonic mean. Precision, recall, and F-scores all range between 0 and 1 (though sometimes this is represented as a percentage between 0 and 100), with higher values indicating a better match to the gold standard.

$$F\text{-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8.11)$$

8.4.2.2 Model Training and Parameter Estimation Because the algorithms used for inference are probabilistic in nature, each modeled strategy (DP-Uni and DP-Bi) was trained and evaluated five times, with the results averaged. Although the learning process modeled is unsupervised, Phillips and Pearl (2014a, 2014b) nonetheless separated the data into training and test sets to better determine how each modeled strategy adapted to new data. Each corpus was randomly split five times so that the training set consisted of 90% of the corpus and the remaining 10% became the corresponding test set. The corpora themselves are temporally ordered, so utterance order captures the order an infant might encounter the utterances. This relative ordering was preserved in both the training and test sets.

The free parameters for the DP-Uni (α) and DP-Bi strategies (β, γ) were set by searching ranges derived from those explored in Goldwater et al. (2009) and Pearl et al. (2011): $\alpha, \beta \in [1, 500]$, $\gamma \in [1, 3000]$. For each language and each strategy, a learner using the idealized inference algorithm was used to

Table 8.2 *Best free parameter values for all unigram and bigram Bayesian segmentation strategies across each language.*

	DP-Uni		DP-Bi	
	α	β	γ	
English	1	1	90	
Italian	1	1	90	
German	1	1	100	
Spanish	1	200	50	
Japanese	1	300	100	
Farsi	1	200	500	
Hungarian	1	300	500	

determine which free parameter values resulted in the best word token F-score. The values identified by this process are shown in Table 8.2.

When we look at the best parameter values, it turns out that the DP-Uni strategy fares best on all languages when $\alpha = 1$. This indicates a strong bias for small proto-lexicons, since novel words are dispreferred. In contrast, the DP-Bi strategy has more variation, roughly breaking into three classes. The first class, represented by English, Italian, and German, has $\beta = 1$ and γ between 90 and 100. This indicates a very strong bias for small proto-lexicons, since novel bigrams are strongly dispreferred (β) and novel words as the second word in a bigram are somewhat dispreferred (γ). The second class, represented by Spanish and Japanese, has β between 200 and 300 and γ between 50 and 100. This indicates a weaker bias for small proto-lexicons, since novel bigrams are only somewhat dispreferred (β) and novel words as the second word in a bigram are also only somewhat dispreferred (γ). The third class, represented by Farsi and Hungarian, has β between 200 and 300 and $\gamma = 500$. This indicates an even weaker bias for small proto-lexicons, since novel bigrams are again only somewhat dispreferred (β) and novel words as the second word in a bigram are even less dispreferred (γ).

Since we are setting these parameter values for our analyses, this translates to the modeled infant already knowing the appropriate values for each language. For the DP-Uni strategy, this may reflect a language-independent bias, since the values are all the same. However, for the DP-Bi strategy, the infant would need to adjust the values based on the ambient language. For either strategy, one potential way to converge on the best values is to have hyperparameters on them and simply learn their values at the same time as segmentation is learned. Incorporating hyperparameter inference into the DP segmentation strategy would be

Table 8.3 *Word token F-scores for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.*

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	Idealized	53.1	60.3	55.0	61.9	66.6	59.9	63.2
	Constrained	55.1	60.3	56.1	58.6	59.6	54.5	63.7
DP-Bi	Idealized	77.1	73.1	64.8	71.3	69.6	66.2	66.5
	Constrained	86.3	82.6	60.2	60.9	62.5	59.5	63.3
Baseline	RandOracle	56.4	47.5	27.0	22.8	20.3	26.4	26.1

a welcome avenue for future segmentation modeling work, particularly if it turns out infants are using something like the DP-Bi strategy. Here we discuss the results obtained by manually setting the free parameters to their respective optimized values in each language.

8.4.2.3 Baseline Strategy: Random Oracle A baseline strategy first explored by Lignos (2012) is a random oracle strategy (RandOracle). The random aspect refers to the strategy treating each possible boundary location as a Bernoulli trial. The oracle aspect comes from the strategy already knowing the true probability of a boundary occurring in the corpus (B Prob in Table 8.1). Boundaries are then randomly inserted with this probability.

8.4.2.4 Cross-Linguistic Performance Table 8.3 presents the gold standard word token F-score results for each learner on all seven languages. First, we can see that bigram assumption is generally helpful, though the degree of helpfulness varies cross-linguistically. For example, it seems very helpful in English and German (e.g., English Idealized DP-Uni: 53.1 vs. DP-Bi: 77.1; German Constrained DP-Uni: 60.3 vs. DP-Bi: 82.6) and not particularly helpful at all in Japanese (Japanese Idealized DP-Uni: 63.2 vs. DP-Bi: 66.5; Japanese Constrained DP-Uni: 63.7 vs. DP-Bi: 63.3). Still, with the exception of the English DP-Uni learner, every single Bayesian learner does better than the random oracle baseline. Interestingly, in English, the DP-Bi constrained learner has the highest score of all learners in all languages. Altogether, this suggests that the DP segmentation strategy is generally a very good one for identifying words in fluent speech.

Nonetheless, something seems to be going on cross-linguistically. Why do we see such variability in segmentation performance (DP-Uni: 53.1–66.6; DP-Bi: 59.5–86.3; RandOracle: 20.3–56.4)? The variability in the random oracle baseline is particularly suggestive that there is something about the languages

themselves, rather than something specific to the DP segmentation strategy. More specifically, English and German seem inherently easier to segment than the other languages.

Fourtassi et al. (2013) suggested that some languages are inherently more ambiguous with respect to segmentation than others. Specifically, even if all the words of the language are already known, some utterances can *still* be segmented in multiple ways (e.g., /gɹ.ejtʃfəl/ segmented as *great full* and *grateful* in English). The degree to which this occurs varies by language, with the idea that languages with high inherent ambiguity are harder to correctly segment. If this is true, we might expect that low inherent segmentation ambiguity correlates to high performance by segmentation strategies. With this in mind, perhaps English and German have lower inherent segmentation ambiguity than the other languages (RandOracle English: 56.4, German: 47.5, Other languages: 20.3–26.4).

In order to quantify this ambiguity, Fourtassi et al. (2013) proposed the normalized-segmentation entropy (NSE) metric:

$$NSE = - \sum_s P_s \log_2(P_s)/(N - 1) \quad (8.12)$$

Here, P_s represents the probability of a possible segmentation s of an utterance and N represents the length of that utterance in terms of potential word boundaries (which is determined by the number of syllables for our learners). To calculate the probability of an utterance, we use the unigram or bigram DP generative model equations described in Section 8.3, since these represent the probability of generating that utterance under a unigram or bigram assumption. As an example, to calculate the NSE of a single utterance /aimgɹ.ejtʃfəl/, we use the unigram and bigram model equations to generate the probability of every segmentation comprised of true English words (P_s above). In this case, two segmentations are possible: *I'm grateful* and *I'm great full*. The probabilities for each segmentation are then used in Equation 8.12 above, with $N = 2$ since there are two potential word boundaries among the three syllables.

Because a low NSE represents a true segmentation that is less inherently ambiguous for the learners using the n-gram assumptions tested here, English and German should have lower NSE scores if inherent segmentation ambiguity was the explanation for the better segmentation performance. Table 8.4 shows the NSE scores for both unigram and bigram learners for all seven languages, with token F-scores for the respective idealized inference learners for comparison.

From Table 8.4, we see that German fits with the hypothesis that low NSE predicts higher segmentation performance, having in both cases the lowest NSE scores. Yet English does not fit this pattern, ranking third/fourth overall for a unigram DP learner and fourth overall for a bigram DP learner. This is despite

Table 8.4 Average NSE scores across all utterances in a language's corpus, ordered from lowest to highest NSE and compared against the idealized inference token F-score for the language. Results are shown for both the DP-Uni and DP-Bi models. Lower NSE scores represent less inherent segmentation ambiguity and higher token F-scores indicate a better segmentation performance.

DP-Uni	NSE	F-score	DP-Bi	NSE	F-score
German	0.000257	60.3	German	0.000502	73.0
Italian	0.000348	61.9	Italian	0.000604	71.3
Hungarian	0.000424	59.9	Hungarian	0.000694	66.2
English	0.000424	53.1	English	0.000907	77.1
Farsi	0.000602	66.6	Spanish	0.00103	64.8
Japanese	0.00126	55.0	Farsi	0.00111	69.6
Spanish	0.00128	63.2	Japanese	0.00239	66.5

English having the lowest token F-scores for the DP-Uni learner and highest token F-scores for the DP-Bi learner. Because of this, the high segmentation performance on both German and English cannot simply be due to both having lower inherent segmentation ambiguity.

More generally, it becomes clear by looking at all seven languages that low NSE does not always lead to higher token F-scores. If it did, we would expect to find a significant negative correlation between NSE score and token F-score – but this does not happen (DP-Uni: $r = -0.084$, $p = 0.86$; DP-Bi, $r = -0.341$, $p = 0.45$). Examining individual languages in Table 8.4, this lack of correlation is apparent. The DP-Uni Farsi NSE score is ranked fifth lowest, but in fact has the highest F-score, while the DP-Uni Spanish NSE score is actually the worst, though it has the second best F-score. When we turn to the DP-Bi learners, we see that Hungarian has the third best NSE score but the next to worst F-score, while English has the fourth worst NSE score but the best F-score. So, NSE is not the principal factor determining segmentation performance, though it may play some role.

An alternative factor comes from considering how often words of the language tend to be monosyllabic. This is captured by the boundary probability in Table 8.1, where English and German both have a much higher probability of having a boundary appear after a syllable (76.26% and 68.60%, respectively, compared to the next highest language Spanish, with boundary probability 53.93%). These are precisely the languages that especially benefit – though only for the DP-Bi and RandOracle learners.

One possible explanation for why boundary probability especially impacts these learners relates to the types of errors these learners make. More

Table 8.5 *Percentage of errors that resulted in an oversegmentation as compared to adult orthographic segmentation.*

		Oversegmentation Errors (%)						
		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	Idealized	1.7	9.1	8.7	39.9	47.7	45.3	39.0
	Constrained	9.0	15.9	25.8	53.8	68.0	55.3	53.5
DP-Bi	Idealized	13.8	26.0	33.0	73.1	59.8	58.0	58.4
	Constrained	44.8	60.6	72.8	89.9	93.4	82.7	79.9
Baseline	RandOracle	51.7	60.0	57.7	57.7	58.7	56.9	54.5

specifically, if a learner tends to oversegment (i.e., incorrectly insert word boundaries), languages that tend to be more monosyllabic already may benefit more – this bias to insert word boundaries may yield true words more often by sheer luck in these languages. We can get a sense of whether a strategy tends to oversegment by looking at the errors it produces. Table 8.5 shows the percentage of all segmentation errors that were oversegmentations for each learner, where possible error types are undersegmentations like *thekitty*, oversegmentations like *the ki tty*, and other segmentation errors like *theki tty*.

If we look at the DP-Uni learners, we see that there *isn't* an oversegmentation tendency for English and German, though there is for most of the rest of the languages. Because English and German are the only two languages where an oversegmentation tendency is specifically beneficial (because they tend to have more monosyllabic words), this may be why the DP-Uni learners don't do much better on English and German compared with the rest of the languages. This contrasts noticeably with the DP-Bi and RandOracle oversegmentation tendencies, which are comparatively much higher for English and German (e.g., DP-Bi Constrained English: 44.8% and RandOracle English: 51.7% vs. DP-Uni Constrained English: 9.0%). So, English and German, which tend to have more word boundaries per utterance anyway, yield better performance for learners that have a stronger tendency to insert word boundaries.

This makes an interesting testable prediction about infant segmentation more generally. If the segmentation strategy infants use leads them to oversegment more often (such as the DP-Bi strategy here, particularly when coupled with constrained inference), we might expect infant segmentations to better match adultlike segmentations in languages whose child-directed speech contains more monosyllabic words (e.g., English and German). In contrast, for languages with fewer monosyllabic words, we would expect infant segmentation to match adultlike segmentation less well.

Table 8.6 *Examples of reasonable errors (with English glosses) made in different languages.* True words refer to the segmentation in the original corpus, while Segmented output represents the segmentation generated by a modeled learner.

		True	Segmented
Real words	Spa	<i>porque</i> “because”	<i>por que</i> “why”
	Jap	<i>moshimoshi</i> “hello”	<i>moshi moshi</i> “if if”
Morphology	Ita	<i>devi</i> “you must”	<i>dev i</i> “must” 2 nd -PL
	Far	<i>miduni</i> “you know”	<i>mi dun i</i> PRES “know” 2 nd -SG
Function words	Ita	<i>a me</i> “to me”	<i>ame</i> “to-me”
	Far	<i>mæn hæm</i> “me too”	<i>mæn hæm</i> “me-too”

8.4.3 A More Flexible Metric: Reasonable Errors

We know that infant segmentation certainly is not a perfect match to adult-like segmentation, given the available observational and behavioral data. With this in mind, Phillips and Pearl (2014a, 2014b, 2015b) considered an output evaluation that allowed the following “reasonable errors” as legitimate early segmentations:

1. Oversegmentations that result in real words (e.g., *grateful* /gɹeɪtʃ/ segmented as *great* /gɹeɪt/ and *full* /fʌl/)
2. Oversegmentations that result in productive morphology (e.g., segmenting off *-ing* /ɪŋ/)
3. Undersegmentations that produce function word collocations (e.g., segmenting *that a* as *that a*)

Table 8.6 offers some examples of each reasonable error type in different languages, while Table 8.7 shows how frequently each error type is made by the modeled learners.

For the errors resulting in at least one true word, Phillips and Pearl (2014a, 2014b, 2015b) included only those occurring at least five times in the corpus because of the reason these types of errors are helpful – namely, they boost the perceived frequency of the true word. For example, if a model

Table 8.7 *Percentage of model errors that produced reasonable errors of different kinds. Real Word Errors represent true words in the corpus occurring at least five times. Morphology Errors represent morphological affixes occurring in the correct location (e.g., suffixes after the main word, prefixes at the beginning). Function Word Errors represent collocations of function words.*

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
Real Word Errors (%)								
DP-Uni	Idealized	1.0	3.3	2.8	23.7	20.1	11.9	17.7
	Constrained	3.4	4.5	6.3	27.6	25.5	14.9	21.4
DP-Bi	Idealized	5.8	7.9	11.2	38.3	24.6	17.8	26.7
	Constrained	29.6	17.6	15.1	57.0	41.5	27.7	34.6
Baseline	RandOracle	17.5	7.7	13.6	14.9	10.0	8.6	12.6
Morphology Errors (%)								
DP-Uni	Idealized	0.2	2.7	2.8	3.3	5.0	2.5	9.1
	Constrained	0.6	4.6	7.5	4.8	7.5	3.4	10.5
DP-Bi	Idealized	1.0	7.7	10.4	6.3	8.4	3.3	10.4
	Constrained	2.6	24.9	20.4	6.7	13.0	4.6	16.9
Baseline	RandOracle	2.2	12.1	10.6	3.0	5.1	3.0	10.1
Function Word Errors (%)								
DP-Uni	Idealized	8.8	27.2	8.9	6.4	4.3	2.3	6.9
	Constrained	10.2	26.7	7.7	5.8	3.1	2.3	7.7
DP-Bi	Idealized	15.7	28.2	6.4	5.8	3.7	2.9	5.2
	Constrained	9.9	10.8	2.3	1.1	0.2	1.7	2.6
Baseline	RandOracle	3.6	8.7	4.2	3.4	1.0	1.1	2.1

segments *grateful* as *great* and *full*, then the next time the word *great* is encountered, it is more likely to be segmented because it has been previously seen. So, only true word errors occurring with some frequency are likely to have this beneficial effect. Additionally, this ensures that typos and nonsense words in the corpus are unlikely to be treated as true words.

We can see in Table 8.7 that certain modeled learners tend to produce more real word errors: learners using constrained inference and learners in Italian, Farsi, and Japanese. Additionally, the DP-Bi learners tend to yield more than the DP-Uni learners. This is likely due to the oversegmentation tendency – these are the same learners that also tend to oversegment more often. Because real word errors occur due to oversegmentation, this makes this error type more likely to occur for learners that oversegment.

For the errors resulting in productive morphology, Phillips and Pearl (2014a, 2014b, 2015b) referenced lists of morphemes for each language produced by linguistically trained native speakers. Similar to the true word errors,

oversegmentations are more likely to produce productive morphology because morphological affixes smaller than words can only be produced through over-segmenting. However, we do note that subsyllabic morphology was excluded, due to the modeled learners perceiving the speech stream as a sequence of atomic syllables. This ruled out much of the common inflectional morphology in Indo-European languages (e.g., *-s* in English). Still, German, Spanish, Farsi, and Japanese have a number of these errors, regardless of the specific modeled learner.

Function word collocation errors, in contrast to the other two types, occur due to undersegmentation. So, learners with a stronger oversegmentation bias, like the constrained DP-Bi learner, produce these relatively less frequently. Looking across the languages, we can see that English and German tend to have more of these errors, perhaps because of the frequency of monosyllabic function words. For example, many of these errors are combinations of the form MODAL VERB+PRONOUN (e.g., *can you*), COPULA+PRONOUN (e.g., *are you*), or PREPOSITION+DETERMINER (e.g., *in a*). The other languages tend to also have richer morphology that can negate the need for separate function words.

Table 8.8 shows the evaluation when we consider all three reasonable error types as acceptable segmentation. The main difference (of course) is that the average token F-score is higher than before (e.g., unadjusted DP-Bi average: 68.9; adjusted DP-Bi average: 77.4). Perhaps more interestingly, we can see that the constrained learners perform as well as or better than the idealized bigram learner on most languages. In cases where they perform noticeably below the idealized learner (DP-Uni: Italian, Hungarian; DP-Bi: Spanish, Italian), they do not perform so poorly that we would consider them unsuccessful (e.g., all are above a word token F-score of 60). This indicates that constrained inference does not necessarily hinder this Bayesian segmentation strategy – and in fact

Table 8.8 Adjusted word token F-scores, counting reasonable errors as acceptable output, for learners across English, German, Spanish, Italian, Farsi, Hungarian, and Japanese. Higher token F-scores indicate better performance, with the best score for each language in bold.

		Eng	Ger	Spa	Ita	Far	Hun	Jpn
DP-Uni	Idealized	56.8	71.1	59.2	68.0	72.1	64.8	71.1
	Constrained	60.0	71.3	64.6	65.1	75.5	61.4	72.7
DP-Bi	Idealized	81.5	82.9	74.7	76.8	76.4	72.0	76.3
	Constrained	90.1	88.4	71.5	71.2	75.1	71.4	75.1
Baseline	RandOracle	64.6	59.4	42.4	31.0	31.7	33.1	41.8

may be helpful for some languages – especially once we incorporate a more nuanced standard of segmentation success.

More generally, many errors made by each learner may not be so harmful. For example, they can potentially be useful for later segmentation (real word errors), identifying productive morphology (morphology errors), grammatical categorization (morphology errors), and the early stages of syntactic bootstrapping (morphology errors, function word collocations). So, the output units of this strategy may be useful for the acquisition process, even if they are not the adultlike segmentation. But how do we tell if they really *are* useful for acquisition?

8.5 How Useful Are the Units?

Model output can be measured in two core ways: intrinsically and extrinsically (Galliers & Jones, 1993). Intrinsic evaluations are concerned with the model's direct objective, i.e., the task it was trained for. Intrinsic measures for speech segmentation include comparisons against the gold standard (e.g., F-score), measures of model fit (e.g., log posterior probability), and comparison against behavioral results from experimental setups (e.g., see Frank, Goodman, and Tenenbaum, 2009 and Kolodny, Lotem, and Edelman, 2015). In contrast, extrinsic measures are concerned with how the model output is used for alternative tasks. For speech segmentation, this relates to how the segmented items are used in the overall acquisition process. That is, because the segmented units are used by other processes during acquisition, it makes sense to try to measure their effectiveness for these secondary tasks. Two ways to do this are (1) incorporate the secondary task into the segmentation model (*joint modeling*), and (2) use the output of the segmentation process to perform the secondary task in isolation (*downstream evaluation*).

Joint modeling for acquisition makes sense when we have reason to believe the two tasks are solved at the same time by infants. So, for example, while joint modeling of early segmentation and phonotactics may be possible, experimental evidence suggests that phonotactic learning begins significantly after early segmentation (Bortfeld et al., 2005; Thiessen & Saffran, 2003; Mattys et al., 1999). This makes it seem cognitively implausible to model both processes jointly.

However, for tasks that may be solved simultaneously, the idea is that the two tasks can bootstrap off each other, resulting in better overall performance in both than if the tasks were solved in isolation. Still, this requires the modeler to make assumptions about what the infant knows with respect to how the two tasks relate to each other, so that the connections between them can be leveraged. More specifically, because these assumptions are typically built into the joint model directly, the implication is that infants already know them *a priori*

(either innately or learned very rapidly in the first months of life). Depending on the particular joint relationship, these assumptions may be plausible – or not.

Downstream evaluation, on the other hand, assumes that there is *no* feedback between the first task and the second; the first task happens first, and its output is used as input to the second task. This may be appropriate when there is a known gap between the tasks, such as early segmentation and phonotactic learning. It may also be appropriate if the tasks have some overlap, but there is reason to think infants do not leverage the synergies between the tasks.

In a sense, joint modeling represents a *best case* scenario for infants. The infant knows that the two tasks are connected and knows specifically how they are related, allowing for joint learning to take advantage of the two processes. In contrast, downstream evaluation represents a *worst case* scenario because the infant does not realize the two tasks are connected and so cannot leverage the additional information available. In reality, infants may often fall somewhere in between these two endpoints, depending on the specific acquisition tasks. By considering both perspectives, we can set upper and lower bounds on what acquisition success looks like. This is particularly helpful for tasks happening early in acquisition, because we are often unsure exactly what infant knowledge representations look like. So, instead of relying only on intrinsic measures that assume these representations take a certain form, we can also use extrinsic measures to evaluate the utility of the inferred representations.

8.5.1 Learning the Right Stress-Based Segmentation Cue

In languages with lexical stress, where stress is placed on syllables within a word, stressed syllables can provide a cue to word boundaries in fluent speech particularly when stressed syllables reliably occur at word edges. In languages with fixed lexical stress, such as Hungarian, this cue is essentially deterministic: every Hungarian word has stress on the initial syllable. In languages with variable lexical stress, such as English, this cue can be probabilistic: the exact position of stressed syllables varies from word to word (e.g., *ápple* vs. *banána*), though there is a reliable tendency for words in English child-directed speech to begin with stressed syllables (Pearl et al., 2011; Phillips & Pearl, 2015b). Importantly, the existence of these reliable stress cues and their exact implementation depend on the language. Given this, there's been significant interest about when infants learn to identify and leverage these stress-based cues to speech segmentation (e.g., Jusczyk et al., 1993; Jusczyk, Houston, & Newsome, 1999; E. Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003; Thiessen & Saffran, 2007).

Because these stress-based cues are language-specific, it is usually thought that infants have to first identify enough words in the language to determine the specific instantiation of the stress-based cue (assuming there is one). The

earliest evidence for infant sensitivity to stress cues is at 7.5 months (Jusczyk, Houston, & Newsome, 1999) while the early stages of segmentation begin around 6 months (Bortfeld et al., 2005). However, given how close in time these two processes occur (stress-cue identification and early segmentation), it is quite possible they overlap. That is, infants could be learning to segment using statistical cues while simultaneously trying to identify if there is a dominant stress pattern in the segmented words. So, a joint model of acquisition seems reasonable. Of course, it is also possible that in reality early segmentation provides a seed pool of segmented words very quickly, and infants subsequently use these to infer a stress-based segmentation cue. So, a downstream evaluation also seems reasonable. We discuss studies using each as extrinsic evaluation metrics for the DP segmentation strategy.

8.5.1.1 Stress Cue Identification: Joint Modeling Doyle and Levy (2013) investigate a joint model of the DP-Bi segmentation strategy (the best-performing DP segmentation variant) and stress pattern identification, using idealized inference. In particular, this model assumes that the learner knows lexical stress exists in the language and attempts to identify the dominant pattern for every sequence of M syllables (i.e., it assumes all two-syllable words have a dominant pattern, and this pattern may be different from the one all three-syllable words have, and so on). This is accomplished via an update to the P_0 probability from (8.3), as shown in (8.13), which now depends both on the probability of the syllables in the word $w_i(P_W)$ and the stress pattern s_i observed for a word with M syllables (P_S).

$$P_0(w_i, s_i) = P(w_i)P_S(s_i|M) \quad (8.13)$$

More specifically, P_W is calculated using P_0 from the original DP-Bi implementation. P_S is calculated as a multinomial over all possible stress patterns of length M (given a uniform prior), with the parameter values derived from observed frequency and plus-one smoothing. This places minimal restrictions on stress patterns for words: a word might possess multiple stressed syllables or no stressed syllables at all.

Doyle and Levy (2013) evaluated their joint model on the Korman corpus of English child-directed speech (Korman, 1984), as modified by Christiansen, Allen, and Seidenberg (1998). This sample contains speech directed at infants between 1.5 and 4 months old, with 24,493 word tokens. Notably, this corpus highlights the monosyllabic bias in English child-directed speech, as the corpus consists of 87.3% monosyllabic words. Phonemic forms, syllabification, and stress patterns were all derived from the MRC Psycholinguistic Database (Wilson, 1988), and demonstrated that this English speech sample had a strong word-initial stress bias (89.2% of all multisyllabic tokens had stress on the first syllable).

Doyle and Levy (2013) compared modeled learners using a joint DP-Bi+Stress strategy against learners using the original DP-Bi strategy. From the perspective of extrinsic measures of segmentation, the question is whether there is bootstrapping observed for both processes. If so, this means the segmented units provide useful information for inferring stress-based cues, and stress-based cues provide useful information for segmentation.

To determine whether the segmented units are helpful for inferring stress-based cues, Doyle and Levy (2013) compared the word-initial stress bias in the segmented words generated by both strategies. Given that the English stress-based cue is that words begin with stressed syllables, modeled learners who succeed should have a bias for stress-initial words over stress-final words. It turned out that while both strategies generated this bias for the segmented units (as shown in Table 8.9), the joint DP-Bi+Stress strategy did so slightly more strongly (DP-Bi+Stress: 87.3% vs. DP-Bi: 86.4% word-initial stress on bisyllabic words). So, there is some synergistic value in the segmented units for the learner – they are useful for more easily inferring the correct stress-based cue in English, with the idea that a stronger bias in the correct direction makes inference easier.

To determine whether the stress-based cues are helpful for segmentation, Doyle and Levy (2013) compared the word token and word type F-scores for both strategies. If there is useful segmentation information contained in the developing representation of the English-specific stress cue, this should be reflected in the F-scores of the DP-Bi+Stress being higher than the DP-Bi alone. As Table 8.9 shows, this does indeed occur (though again, the benefit is

Table 8.9 Extrinsic measure evaluations for the joint model from Doyle & Levy (2013) (DP-Bi+Stress) compared against the original DP-Bi model. Stress bias indicates the bias toward the correct English stress-based cue to segmentation, with higher percentages indicating a stronger bias. F-scores are shown either including frequency (Word token) or factoring it out (Word type). Higher scores indicate better segmentation performance compared against the gold standard.

	DP-Bi+Stress	DP-Bi
Stress bias	87.3%	86.4%
Word token F-score	68	67
Word type F-score	80	77

somewhat slight, given the excellent segmentation performance the DP-Bi already achieves on its own).

These results indicate that jointly inferring boundaries and stress cues does yield a bootstrapping effect for both tasks. However, because this effect is rather small, it may well be that infants can do just fine even if these processes are not occurring simultaneously.

8.5.1.2 Stress Cue Identification: Downstream Evaluation Here we consider the downstream evaluation for the DP segmentation strategy explored by Phillips and Pearl (2015a), where the segmentation process yields a proto-lexicon. We can then evaluate that proto-lexicon with respect to inferring the correct stress-based cue. Phillips and Pearl (2015a) did this for several DP segmentation variants, some of which had significantly lower F-score performance than others. This can give us a sense of how good segmentation needs to be in order for the units in the proto-lexicon to be helpful for inferring the correct stress-based cue.

The evaluation process itself is similar to the approach taken by Doyle and Levy (2013): examine the proto-lexicon yielded by different DP segmentation strategy variants and calculate the bias toward word-initial stress for each. Table 8.10 shows the adjusted segmentation F-scores achieved by each strategy against the gold standard on English child-directed speech from the Brent corpus, when reasonable errors are counted as acceptable. It additionally shows the strength of the bias in the inferred proto-lexicon toward word-initial stress. To

Table 8.10 Downstream evaluation for different DP segmentation strategy variants, compared against the adult orthographic segmentation and the random oracle baseline in English, German, and Hungarian. Adjusted word token F-scores are shown, which count reasonable errors as acceptable segmentation. The percentage of bisyllabic word types in the inferred proto-lexicon with word-initial stress are shown, given all bisyllabic word types the learner identified containing a single stressed syllable.

		Adjusted F-score			% Stress-initial items		
		English	German	Hungarian	English	German	Hungarian
Unigram	Adult seg	100	100	100	88.4%	90.3%	100%
	Idealized	56.8	71.1	64.8	87.3%	90.8%	96.9%
	Constrained	60.0	71.3	61.4	85.3%	87.6%	93.1%
Bigram	Idealized	81.5	82.9	72.0	88.4%	90.9%	97.9%
	Constrained	90.1	88.4	71.4	90.6%	92.6%	96.4%
Baseline	RandOracle	64.6	59.4	33.1	46.7%	49.6%	52.5%

determine the stress patterns for segmented words, Phillips and Pearl (2015a) referenced the English Callhome Lexicon (Kingsbury, Strassel, McLemore, & MacIntyre, 1997). Child-register words not found in standard dictionaries (e.g., *moosha*) were manually coded when the proper stress could be reasonably inferred. Additionally, to better approximate the stress of words in fluent speech, monosyllabic words were left unstressed. Similar stress analyses were done for German and Hungarian, using the Caroline (German) and Gervain (Hungarian) corpora of child-directed speech, and determining stress patterns by using the Callhome German Lexicon (Karins, MacIntyre, Brandmair, Lauscher, & McLemore, 1997) and the stress rules of Hungarian (all words are stress-initial).

We can first observe that if segmentation matches adult orthographic segmentation (with an F-score of 100), the stress-initial bias is quite strong: 88.4% (English), 90.3% (German), or 100% (Hungarian) of bisyllabic word types with a single stressed syllable have that syllable at the beginning of the word. This should make inferring the stress-based segmentation cue straightforward. Interestingly, every single DP segmentation learner – regardless of its F-score performance – achieves a stress-initial bias that's almost as strong (English, German, Hungarian), as strong (English, German), or stronger (English, German). That is, even strategies whose segmentation seems poorer (DP-Uni learners) or actually worse than the baseline on F-score (e.g., DP-Uni learners in English) achieve a very strong stress-initial bias in their proto-lexicons. It doesn't matter that their segmentation doesn't match the adult orthographic standard; it is good enough from the perspective of inferring the correct stress cue to segmentation.

This contrasts notably with the baseline random oracle strategy, which is the only strategy to identify a proto-lexicon that fails to have a bias (German), has only a very slight bias in the correct direction (Hungarian), or actually has a bias in the wrong direction (English, where less than 50% of its bisyllabic word types are stress-initial). This occurs even though its adjusted English F-score is higher than the F-scores of the DP-Uni learners. This underscores that even if a strategy's output doesn't match adult segmentation, that output can still be quite useful, as we see here especially with the DP-Uni learners. The important stress pattern property is preserved in the inferred proto-lexicon.

8.5.1.3 Stress Cue Identification: Summary Both the joint modeling and downstream extrinsic metrics suggest that the DP segmentation strategy is quite useful for identifying the stress-based segmentation cue for a language. It is particularly notable that this can occur for the downstream evaluation even if the segmentation does not match the adult segmentation very well, as indicated by the standard metric of the F-score. So, even lower-quality proto-lexicons may be good enough for subsequent acquisition processes such

as identifying the language-specific stress cue to speech segmentation. We turn next to another acquisition process that relies on the output of segmentation.

8.5.2 Learning the Meaning of Concrete Nouns

Infants begin associating word forms from their proto-lexicons with concrete objects such as *cookie* and *nose* when they are between six and nine months old (Bergelson & Swingley, 2012), and so this process could very well overlap with early speech segmentation. Frank et al. (2009) proposed a Bayesian word-mapping strategy for concrete objects that incorporates an infant's developing ideas about referential intention, which they called the Intentional strategy. The Intentional strategy identifies a comparatively accurate lexicon of word-meaning mappings, and accounts for several well-known phenomena in the developmental word-mapping literature, including cross-situational word learning (Yu & Smith, 2007; Smith & Yu, 2008; Yu & Smith, 2011), mutual exclusivity (Markman & Wachtel, 1988; Markman, 1989; Markman, Wasow, & Hansen, 2003), one-trial learning (Carey, 1978; Markson & Bloom, 1997), object individuation (Xu, 2002), and intention reading (Baldwin, 1993). One assumption Frank et al.'s (2009) implementation makes is that speech is represented in its adult orthographic segmentation. Given the Intentional strategy's success at matching developmental data and its reliance on segmented speech, it seems reasonable to use it either within a joint model of segmentation and word learning or as a downstream evaluation of a segmentation strategy's output. We first give an overview of the Intentional strategy and then discuss each evaluation in turn.

8.5.2.1 The Intentional Word Learning Strategy Because the Intentional strategy is a Bayesian strategy that relies on a generative model, its components can be represented with a plate diagram as shown in Figure 8.1. The modeled infant using this strategy observes the individual words uttered (W_s) in a particular situation s and the concrete objects O_s in the vicinity. The learner then infers which objects I_s the speaker intends to refer to as well as the lexicon L the speaker is drawing from in order to make the words W_s refer to those objects. That is, the lexicon L is a set of word-meaning mappings, drawn from the word forms W and the available objects O in all observed situations.

The probability of a corpus of situations S , given an adult speaker's lexicon L , can be represented as in (8.14). It is the product of the probability of each individual situation s , where the probability of each possible intended object I_s is summed (because the actual intention of the speaker is not observed).

$$P(C|L) = \prod_{s \in S} \sum_{I_s \in O_s} P(I_s|O_s)P(W_s|I_s, L) \quad (8.14)$$

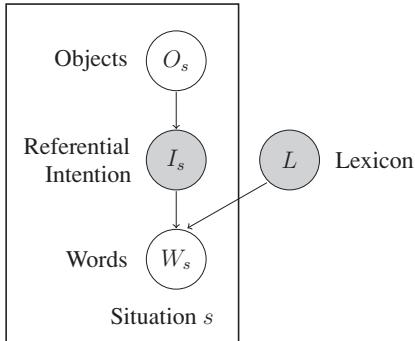


Figure 8.1 Plate diagram of the Intentional strategy's generative model.

The probability of intending to speak about any available object, $P(I_s|O_s)$, is treated as uniform, such that all objects are equally likely to be referred to. The difference between intended objects comes from the probability of selecting the words W_s given the speaker's intentions in that situation I_s and the speaker's lexicon ($P(W_s|I_s, L)$). The modeled learner is aware that some words may be spoken nonreferentially (i.e., they do not map to a concrete object), and so this probability is calculated as in (8.15). More specifically, there is some probability γ that the word is used referentially and some probability $(1 - \gamma)$ that it is not. The best-performing variant of the Intentional strategy reported in Frank et al. (2009) used $\gamma = 0.1$, implementing a strong bias for words to be used nonreferentially. This makes intuitive sense as only a few words (and often only a single word) in an utterance actually refers to a concrete noun (e.g., In *Look at the kitty!*, only *kitty* is referential in this sense).

$$P(W_s|I_s, L) = \prod_{w \in W_s} [\gamma \sum_{o \in I_s} \frac{1}{|I_s|} P_R(w|o, L) + (1 - \gamma) P_{NR}(w|L)] \quad (8.15)$$

The probability for all observed words W_s is the product of each individual word w . If a word is used referentially, the learner assumes it's chosen uniformly from all words linked in the lexicon to the intended object in question ($P_R(w|o, L)$). For instance, if *duck* is the only word linked to the object DUCK, then it has probability 1. If both *duck* and *bird* are linked to the object DUCK, then each has probability 0.5. This probability is summed across all potential intended objects I_s and averaged.

In contrast, if a word is used nonreferentially (in the sense that it does not refer to an available concrete noun in O_s), the learner then distinguishes within $P_{NR}(w|L)$ between words that are in the lexicon and words that are not. A word can be used nonreferentially even if it is in the lexicon because a word form may map to more than one meaning (e.g., consider the concrete noun *duck* vs.

the action verb *duck*), and the learner allows this possibility. So, if a word in the lexicon is used nonreferentially, it is selected with probability proportional to κ . In contrast, if a word not in the lexicon is used nonreferentially, it is selected with probability proportional to 1. If $\kappa < 1$, words with entries in the lexicon are less likely to be used nonreferentially than words that are not. The best-performing variant of the Intentional strategy reported in Frank et al. (2009) used $\kappa = 0.05$, implementing a strong bias for words in the lexicon not to be used nonreferentially. That is, if a word has an entry in the lexicon mapping the word form to one or more concrete objects, this learner strongly prefers the word to be used to refer to a concrete object.

8.5.2.2 Learning Concrete Nouns: Joint Modeling M. Johnson et al. (2010) explore one approach to a joint model of segmentation and word-object mapping. They implement these models using Adaptor Grammars (AGs), an extension of probabilistic context free grammars (PCFGs) that can be used to model segmentation. Using AGs, they implement both the DP-Uni and DP-Bi segmentation strategies, as well as joint models that incorporate the Intentional strategy for learning a concrete noun lexicon. They implement two variants of the Intentional strategy: the original version (+Intention), which they refer to as *reference*, and one that builds in a hard constraint that a word can only refer to a single concrete object within an utterance (+Intention+Only1), which they refer to as *reference1* (for DP-Uni) and *referenceC1* (for DP-Bi). In effect, in (8.15), the +Intention+Only1 learner has an additional constraint on P_R that considers whether the word has already been used referentially for an intended object in I_s . If it has, $P_R = 0$.

M. Johnson et al. (2010) used idealized inference and evaluated these strategies on the Fernald-Morikawa corpus (Fernald & Morikawa, 1993), which consists of 22,000 words (5,600 utterances) of mother-child play sessions involving pairs of toys. These utterances were phonemically transcribed using the VoxForge dictionary and segmentation assumed phonemes were the basic units of representation. Given the empirical data about infant speech representation, it turns out this may not be the most plausible assumption. Nonetheless, M. Johnson et al. (2010) found synergies between speech segmentation and concrete object learning, as shown in Table 8.11, and so future studies may wish to replicate these investigations using syllables as the basic unit of representation.

First, we can ask if knowing that some of the units in speech refer to available concrete objects improves segmentation. Whether the modeled learner is using the DP-Uni or DP-Bi strategy variant, the answer is clearly yes, though the improvement is more substantial for the DP-Bi learner (e.g., DP-Uni Base = 53.3 vs. +Intention+Only1 = 54.7; DP-Bi Base = 69.5 vs. +Intention+Only1 = 75.0). Moreover, hard-wiring in the constraint that words within an utterance can only refer to at most a single concrete object is helpful (DP-Uni

Table 8.11 *Extrinsic measure evaluations for the joint model (+Intention, +Intention+Only1) from Johnson et al. (2010) compared against the original DP segmentation strategies in isolation (Base). Segmentation F-score indicates how well the modeled learner segmented the utterances when compared to the adult orthographic gold standard. Lexicon F-score indicates how well the modeled learner identified the word-object mappings. For both F-scores, higher scores indicate better performance.*

		Segmentation F	Lexicon F
DP-Uni	Base	53.3	0.0
	+Intention	53.7	14.9
	+Intention+Only1	54.7	14.7
DP-Bi	Base	69.5	0.0
	+Intention	72.6	22.0
	+Intention+Only1	75.0	63.6

+Intention = 53.7 vs. +Intention+Only 1 = 54.7; DP-Bi +Intention = 72.6 vs. +Intention+Only 1 = 75.0), though segmentation performance is already quite good for all DP segmentation strategy variants.

Second, we can ask whether knowing the segmented units improves identification of the mappings from word forms to concrete objects in the lexicon. M. Johnson et al. (2010) used a default assumption that no words in the utterance are referential, and so the baseline performance for the DP-Uni and DP-Bi segmentation strategies alone is 0.0. As a point of reference, Frank et al. (2009) achieved a lexicon F-score of 55.0 on the corpus they used, and this was twice as good as the next closest strategy, which had a lexicon F-score of 22.0. Looking at the joint model results, it seems that only the DP-Bi joint model with the constraint restricting a word form's referent within an utterance (DP-Bi+Intention+Only1) achieves a noteworthy lexicon F-score (63.6). Still, it is higher than Frank et al.'s (2009) previous results that used the adult orthographic segmentation (though again, we note that was on a different corpus). This suggests that the imperfectly segmented units are indeed helpful, but only the more sophisticated segmentation strategy (DP-Bi) generates segmentations that are good enough for a joint model to benefit from them, and only if the word-mapping portion of the joint model contains that additional restriction on word reference within an utterance.

More generally, if infants are using joint learning strategies of this kind, these results indicate that the imperfect segmentations generated by the DP segmentation strategy may be sufficient to bootstrap word-object mapping for concrete nouns. In particular, the more sophisticated DP-Bi assumption is the most

Table 8.12 *Segmentation and word-object mapping results for modeled learners in Phillips & Pearl (2015a). Token F-scores are shown for segmentations both on the original Brent corpus and the Rollins corpus. Lexicon precision is shown for the Rollins corpus, representing the accuracy of the word-object mapping model trained on each learner's segmentation.*

		Segmentation		Mapping
		Token F		Lexicon P
		Brent	Rollins	Rollins
DP-Uni	Adult Segmentation	100	100	58.3
	Idealized	53.1	51.4	46.5
DP-Bi	Constrained	55.1	52.4	47.3
	Idealized	77.1	74.6	54.4
Baseline	Constrained	86.3	81.3	38.8
	RandOracle	56.4	57.6	40.6

promising of the DP segmentation variants. Interestingly, this is also supported by the results of the downstream evaluation we discuss in the next section, though whether the learner uses idealized vs. constrained inference turns out to matter.

8.5.2.3 Learning Concrete Nouns: Downstream Evaluation

Phillips and Pearl (2015a) investigated a downstream evaluation of the syllable-based DP segmentation strategy using the Intentional strategy for learning concrete nouns. First, each modeled learner was trained on a subsection of the Brent English corpus of child-directed speech (Brent & Siskind, 2001), which is naturalistic speech directed at children between six and nine months and consists of 28,391 utterances. Using the proto-lexicon it had inferred, each modeled learner then segmented the corpus used by Frank et al. (2009), which is derived from two video files from the Rollins corpus (Rollins, 2003) directed at six-month-olds, where caretakers were asked to play with their infants in an experimental setup. These files were hand-annotated for concrete objects in the immediate vicinity, yielding 619 utterances. The segmented Rollins utterances were then used as input to the Intentional strategy, along with the annotations of available concrete objects in each utterance. The results for both segmentation and inferring the lexicon of concrete nouns are shown in Table 8.12.

The first thing to notice is that segmentation performance transfers quite well between the Brent and Rollins corpora, no matter which learner we look at

(e.g., DP-Bi Idealized: Brent = 77.1 vs. Rollins = 74.6). This highlights the generalizability of the segmentation knowledge each learner has internalized by first encountering the Brent data.

Turning to the lexicon of word-mappings inferred by the learners, Phillips and Pearl (2015a) chose to focus on the precision only, with the idea that it is more important for the very early stages of word learning to find a highly accurate set of mappings, even if not all correct mappings are identified. Moreover, certain principles of word learning infants follow, such as mutual exclusivity (Markman & Wachtel, 1988; Markman et al., 2003), would prevent them from learning all possible lexicon mappings in the Rollins corpus (e.g., both *rabbit* and *bunny* can refer to RABBIT). So, Phillips and Pearl (2015a) reasoned precision was the more relevant metric, rather than the F-score, which additionally incorporates recall.

While the Intentional strategy trained on the adult orthographic segmentation did the best (58.3), all of the DP segmentation strategy variants yielded segmentations the Intentional strategy found more useful than a random segmentation (40.6), except for the constrained DP-Bi learner. Strikingly, the constrained DP-Bi learner yielded the least accurate lexicon (38.8), despite having the highest token F-scores among the modeled learners (Brent: 86.3, Rollins: 81.3). This surprising result underscores the importance of extrinsic evaluation metrics – just because a strategy fares well on intrinsic measures like token F-score does not mean it will be useful, as assessed by extrinsic measures like downstream evaluation.

On the other hand, of course, good performance on an intrinsic measure does not automatically yield poor performance on extrinsic measures. The idealized DP-Bi learner, in contrast to the constrained version, yields the most accurate lexicon (54.4) among the modeled learners and has the second highest segmentation performance on the Rollins corpus (74.6). This is in line with the joint modeling results from M. Johnson et al. (2010), who also found that the best lexicon resulted from the idealized DP-Bi learner. Also, both DP-Uni learners yield better lexicons than the random oracle baseline (DP Uni Idealized = 46.5, Constrained = 47.3), despite having significantly poorer segmentation performance.

8.5.2.4 Learning Concrete Nouns: Summary The results of both the joint and downstream evaluations using word-object mapping suggest that the DP segmentation strategy (particularly the DP-Bi version) yields segmentations that are useful for learning a lexicon of concrete nouns. Like the findings for stress cue identification, it is notable that yielding a less adultlike segmentation does not necessarily mean the segmentation is not useful for word learning. So, as before, even lower-quality proto-lexicons (as measured against

adult orthography) may be good enough for acquisition processes that depend on those segmented units.

8.5.3 *Extrinsic Evaluations: Summary*

More generally, these extrinsic evaluations suggest that judging the quality of an acquisition strategy from multiple perspectives, both intrinsic and extrinsic, is worth doing. Output that may not intrinsically seem so good could very well be good enough to accomplish what it needs to for the language acquisition process. In terms of extrinsic evaluation options, a joint model may be a good option when two tasks are closely interrelated and overlap in development. However, implementing a joint model can prove technically challenging, depending on the modifications required, and requires assumptions about infant learning that may or may not be well-founded. In contrast, downstream evaluation can be applied without altering the learning strategy for either task, though it may underestimate the synergistic information available to children.

8.6 Closing Thoughts

In this chapter, we hope to have shown that the output of early acquisition processes, such as speech segmentation, is not necessarily the knowledge that an adult has. Nonetheless, this output may very well be useful for infants by providing units that scaffold other acquisition processes. Given this goal, the Bayesian segmentation strategy seems effective for all seven languages tested. Moreover, because learners using this segmentation strategy are looking for useful units, which can be realized in different ways across languages, they can identify foundational aspects of a language that are both smaller and larger than orthographic words.

References

- Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology, 29*(5), 832.
- Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258.
- Bertoni, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology, 117*(1), 21–33.
- Best, C., McRoberts, G., LaFleur, R., & Silver-Isenstadt, J. (1995). Divergent developmental patterns for infants' perception of two nonnative consonant contrasts. *Infant Behavior and Development, 18*, 339–350.

- Best, C., McRoberts, G., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 345–360.
- Bijeljac-Babic, R., Bertoni, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology, 29*(4), 711–721.
- Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phono-tactic cues to the problem of word segmentation. *Journal of Child Language, 37*, 487–511.
- Bonawitz, E., Denison, S., Chen, A., Gopnik, A., & Griffiths, T. (2011). A simple sequential algorithm for approximating bayesian inference. In *Proceedings of the 33rd annual conference of the cognitive science society*, 2463–2468.
- Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science, 16*(4), 298–304.
- Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34*, 71–105.
- Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary. *Cognition, 81*, 31–44.
- Brown, R. (1973). *A first language: The early stages*. Harvard University Press.
- Carey, S. (1978). The child as word learner. In J. Bresnan, G. Miller, & M. Halle (Eds), *Linguistic theory and psychological reality*. (pp. 264–293). Cambridge, MA: MIT Press.
- Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*(2–3), 221–268.
- Cole, R., & Jakimik, J. (1980). Perception and production of fluent speech. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163) Hillsdale, NJ: Erlbaum.
- Cornell, E. H., & Bergstrom, L. I. (1983). Serial-position effects in infants' recognition memory. *Memory & Cognition, 11*(5), 494–499.
- Davis, S. J., Newport, E. L., & Aslin, R. N. (2011). Probability-matching in 10-month-old infants. *Proceedings of the 33rd Cognitive Science Society*, 3011–3015.
- Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition, 126*, 285–300.
- Doyle, G., & Levy, R. (2013). Combining multiple information types in Bayesian word segmentation. In *Highlights – North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 117–126).
- Eimas, P. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America, 105*(3), 1901–1911.
- Ferguson, T. (1973). A Bayesian analysis of Some Nonparametric Problems. *Annals of Statistics, 1*(2), 209–230.
- Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development, 64*(3), 637–656.

- Fourtassi, A., Börschinger, B., Johnson, M., & Dupoux, E. (2013). Why is English so easy-to-segment. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics* (pp. 1–10).
- Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20, 579–585.
- Galliers, J., & Jones, K. S. (1993). *Evaluating natural language processing systems*. (Tech. Rept. No. 291). Computer Laboratory, University of Cambridge.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6, 721–741.
- Gervain, J., & Erra, R. G. (2012). The statistical signature of morphosyntax: A study of Hungarian and Italian infant-directed speech. *Cognition*, 125(2), 263–287.
- Goldwater, S., Griffiths, T., & Johnson, M. (2009). A bayesian framework for word segmentation. *Cognition*, 112(1), 21–54.
- Goldwater, S., Griffiths, T., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12, 2335–2382.
- Gulya, M., Rovee-Collier, C., Galluccio, L., & Wilk, A. (1998). Memory processing of a serial list by young infants. *Psychological Science*, 9(4), 303–307.
- Hohne, E., & Jusczyk, P. (1994). Two-month-old infants' sensitivity to allophonic differences. *Perception & Psychophysics*, 56(6), 613–623.
- Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 548–567.
- Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the tenth meeting of the ACL special interest group on computational morphology and phonology* (pp. 20–27).
- Johnson, M., & Demuth, K. (2010). Unsupervised phonemic Chinese word segmentation using adaptor grammars. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 528–536).
- Johnson, M., Demuth, K., Jones, B., & Black, M. J. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems* (pp. 1018–1026).
- Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- Jusczyk, P., Cutler, A., & Redanz, N. (1993). Infants' preference for the predominant stress pattern of English words. *Child Development*, 64(3), 675–687.
- Jusczyk, P., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, 23(5), 648–654.
- Jusczyk, P., Hohne, E., & Baumann, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61, 1465–1476.
- Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39, 159–207.
- Jusczyk, P., Jusczyk, A., Kennedy, L., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, 21(4), 822–836.

- Kam, C. H., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151–195.
- Kam, C. L. H., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59(1), 30–66.
- Karins, K., MacIntyre, R., Brandmair, M., Lauscher, S., & McLemore, C. (1997). *CALL-HOME German Lexicon*. Linguistic Data Consortium.
- Kingsbury, P., Strassel, S., McLemore, C., & MacIntyre, R. (1997). *CALLHOME American English Lexicon (PRONLEX)*. Linguistic Data Consortium.
- Kolodny, O., Lotem, A., & Edelman, S. (2015). Learning a generative probabilistic grammar of experience: A process-level model of language acquisition. *Cognitive Science*, 39, 227–267.
- Köpcke, K.-M. (1998). The acquisition of plural marking in English and German revisited: Schemata versus rules. *Journal of Child Language*, 25(2), 293–319.
- Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language*, 5, 44–45.
- Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255, 606–608.
- Lignos, C. (2012). Infant word segmentation: An incremental, integrated model. In *Proceedings of the 30th West Coast conference on formal linguistics* (pp. 237–247).
- Lignos, C., & Yang, C. (2010). Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 88–97).
- MacWhinney, B. (2000). *The childe project: Tools for analyzing talk*. 3 edn. Mahwah, NJ: Lawrence Erlbaum Associates.
- Markman, E. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.
- Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20, 121–157.
- Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, 47, 241–275.
- Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385, 813–815.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt and Co. Inc.
- Marthi, B., Pasula, H., Russell, S., & Peres, Y. (2002). Decayed MCMC filtering. In *Proceedings of 18th UAI* (pp. 319–326).
- Mattys, S., Jusczyk, P., & Luce, P. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38, 465–494.
- Pearl, L. (2014). Evaluating learning strategy components: Being fair. *Language*, 90(3), e107–e114.
- Pearl, L., Goldwater, S., & Steyvers, M. (2011). Online learning mechanisms for bayesian models of word segmentation. *Research on Language and Computation*, 8(2), 107–132. (special issue on computational models of language acquisition).
- Peters, A. (1983). *The units of language acquisition*. New York: Cambridge University Press.

- Phillips, L. (2015). *The role of empirical evidence in modeling speech segmentation*. Unpublished doctoral dissertation, University of California, Irvine.
- Phillips, L., & Pearl, L. (2012). 'Less is More' in Bayesian word segmentation: When cognitively plausible leaners outperform the ideal. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 863–868).
- Phillips, L., & Pearl, L. (2014a). Bayesian inference as a cross-linguistic word segmentation strategy: Always learning useful things. In *Proceedings of the computational and cognitive models of language acquisition and language processing workshop* (pp. 9–13).
- Phillips, L., & Pearl, L. (2014b). Bayesian inference as a viable cross-linguistic word segmentation strategy: It's all about what's useful. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (p. 2775–2780). Quebec City: Cognitive Science Society.
- Phillips, L., & Pearl, L. (2015a). Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics 2015* (pp. 68–78). NAACL.
- Phillips, L., & Pearl, L. (2015b). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, 39(8), 1824–1854.
- Polka, L., & Werker, J. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2), 421–435.
- Rollins, P. (2003). Caregiver contingent comments and subsequent vocabulary comprehension. *Applied Psycholinguistics*, 24, 221–234.
- Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2001). Visual short-term memory in the first year of life: Capacity and recency effects. *Developmental Psychology*, 37(4), 539–549.
- Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, 17(4), 443–464.
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.
- Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50, 86–132.
- Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.
- Teinonen, T., Fellman, V., Näätänen, R., Alku, P., & Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, 10(1), 21.
- Thiessen, E., & Saffran, J. (2007). Learning to learn: Infant's acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, 3(1), 73–100.
- Thiessen, E., & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, 39(4), 706–716.
- Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2), 172–175.

- Tincoff, R., & Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. *Infancy*, 17(4), 432–444.
- von Luxburg, U., Williamson, R., & Guyon, I. (2011). Clustering: Science or Art? In *JMLR workshop and conference proceedings* 27 (pp. 65–79).
- Werker, J., & Lalonde, C. (1988). Cross-language speech perception: initial capabilities and developmental change. *Developmental Psychology*, 24(5), 672–683.
- Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior & Development*, 7, 49–63.
- Wilson, M. (1988). The MRC psycholinguistic database machine readable dictionary. *Behavioral Research Methods, Instruments and Computers*, 20, 6–11.
- Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3), 223–250.
- Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414–420.
- Yu, C., & Smith, L. B. (2011). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*, 14(2), 165–180.

Part IV

Social and Language Evolution

9 Social Evolution of Public Languages: Between Rousseau's *Eden* and Hobbes' *Leviathan*

Anne Reboul

Abstract

In the present study, I argue for a two-step or dual account of language evolution, in which structural features of language (discrete infinity, semanticity, decoupling) evolved as part of a Language of Thought, which was then exapted for communication. At the second step, given the social nature of communication, a social scenario is needed. I first examine and reject two social scenarios: the change in social organization proposed by Dunbar (an increase in group size in modern humans) and the change in prosocial attitudes advocated by Tomasello (where modern humans evolved toward altruism). I settle for a mildly Machiavellian account (between Rousseau and Hobbes), in which language evolved to allow humans to manipulate each other, though the manipulation, while beneficial to the speaker, is not necessarily detrimental to the hearer. A strong clue to the manipulative nature of linguistic communication lies in implicit communication (presupposition and conversational implicature) that allows speakers not only to hide their manipulative intentions but even to deny that they had such intentions.

9.1 Introduction

Most accounts of language evolution see it as having evolved *for* communication. Let us say that such accounts see language as a *communication system in the strong sense*. These accounts rest on an undeniable and obvious fact: humans in communication routinely use language. The view that language is a communication system in the strong sense, given that communication is the epitome of a social process, has rather understandably led to the idea that language evolution was first and foremost a social phenomenon. Indeed, most extant accounts of language evolution propose “social” scenarios: Számado and Szathmáry (2006) list eleven scenarios (gossip, grooming, group bonding/ritual, hunting, language as a mental tool, pair bonding, sexual selection, song, status for information, and tool making), only one of which – language as a

mental tool – clearly is not “social.” So, there seems to be a fairly large consensus that language is a communication system in the strong sense that it evolved for social reasons. It should be added that the social scenarios enumerated earlier can be divided according to whether they see the social pressure leading to the emergence of language as due to prosocial attitudes (the *cooperative/altruistic* hypothesis advocated by Tomasello 2009, 2010) or to an arm race motivated by inside group competition and conflict (see Dunbar 1996, 1998).¹ The first are on *Rousseau*’s side, seeing humans as fundamentally altruistic. The second are *Hobbesian* in nature, seeing conflict and competition as the basis of human psyche. On the whole, there is a majority of cooperative/altruistic scenarios for language evolution, and Machiavellian scenarios are the exception rather than the rule. So, in a nutshell, an overwhelming majority of accounts of language evolution see language as a communication system in the strong sense, and are based on social scenarios, most of which are cooperative/altruistic.

This raises two basic questions:

- Is language really a communication system in the strong sense?
- If the emergence of language was due to social pressures, were those social pressures cooperative/altruistic or Machiavellian?

I will now discuss both questions in that order.

9.2 Is Language a Communication System in the Strong Sense?

In the preceding section, we saw that most accounts of language evolution view language as a communication system *in the strong sense*, i.e., as having evolved *for* communication. On the face of it, this seems commonsensical: What else could language be, or, in other words, what else could have language evolved *for*? Yet, the idea that language is a communication system in the strong sense meets with fundamental difficulties, due to the deep differences between it and other animal communication systems.

Most accounts of language evolution see language as continuous with other animal communication systems (see, e.g., Millikan 1984, 2004, 2005; Tomasello 2009, 2010; Dunbar 1996, 1998, among many others).² Yet, there are major differences between language and all other animal systems, and it is important to identify these differences for two reasons:

¹ Note that both of these theories of language evolution have counterparts regarding the evolution of cognition (see Tomasello 2014 for a cooperative/altruistic scenario for the evolution of cognition; and the papers in Byrne and Whiten 1988, Whiten and Byrne 1997, as well as Maestripieri 2007 for a popular exposition, for the Machiavellian scenario of the evolution of cognition).

² For a recent and original dissenting view, see Scott-Phillips (2015).

- They are the *explananda* for any theory of language evolution.
- As we shall see, they raise major difficulties for the notion that language is a communication system in the strong sense.

In other words, any theory of language evolution has to explain why humans (and *only* humans) needed a system of communication with these unique characteristics. And any theory of language evolution that claims that language is a communication system in the strong sense has to explain why these characteristics are not an obstacle to this specific claim.

On the face of it, it might seem that there is no characteristic that is specific to language.³ As was shown by Fitch (2010), all of the 13 features listed by Hockett (1963) as characteristic of language can be found in another animal communication system:

- The use of the vocal channel is in fact extremely widespread in animal communication systems, both in birds and mammals; as a consequence, so are broadcast transmission, rapid fading, and total feedback.
- Interchangeability is found in the alarm signals of birds and mammals.
- Specialization is the basis of most animal communication systems.⁴
- Semanticity is again found in the alarm calls of birds and mammals, where some species have different calls for different predators.
- Arbitrariness is found in most animal communication systems.
- Decoupling can be found (as far as is known) only in the honeybee dance through which a bee informs other bees *inside* the hive of the location of nectar *outside* the hive.
- Duality of patterning is found in some oscine birds, who not only learn their songs but produce new repertoire of songs each spring.
- Productivity and discreteness that derive from duality of patterning are also found in these species.
- Traditional transmission can also be found in all oscine birds, as well as in the gestural communication of great apes.

However, there is a core combination of features that seem unique to language in the sense that, although each feature can be found in another animal communication system, the combination as such is found nowhere but in language. This core combination of features gathers *semanticity*, *discrete infinity*, and *decoupling*. It has a major consequence: language is not only unlimited in the number of different sentences it can produce (through *discrete infinity*); these

³ This would basically mean that there is no necessity for a theory of language evolution. A general theory of the evolution of communication would do.

⁴ Indeed, animal signals, being phylogenetically inherited rather than learned, are, if anything, more specialized than anything linguistic can be (see, e.g., Hauser 1996).

sentences also have different contents (through *semanticity*), and these contents are largely independent on the situation the speaker finds oneself in (through *decoupling*). This leads to two major questions that are especially difficult to answer for any account that views language as a communication system in the strong sense:

- Why did humans (and only humans) need a communication system that allows them to produce infinity of different sentences with different contents?⁵
- How did language evolve as a communication system in the strong sense (i.e., *for communication*), given that it incorporates the perfect tool for deception (decoupling)?

The second question is linked to one of the basic tenets of work on the evolution of communication (for a lucid exposition, see Maynard Smith and Harper 2003). For a communication system to evolve, it has to be beneficial to both the sender and the receiver (as argued by Krebs and Dawkins 1984). Deception makes a signal detrimental to the receiver. If deception were widespread, receivers would be selected to ignore signals, which would put an end to the evolution of the communication system. In the overwhelming majority of communication systems,⁶ signals are *not* decoupled. This allows receivers to check the veracity of signals as soon as they are produced, making deception an unprofitable endeavor. However, language *is* decoupled, making it more often than not impossible for the hearer to check whether the content is true or not. In Trivers's (2011, loc. 209) words, "our most prized possession – language – not only strengthens our ability to lie but greatly extends its range. We can lie about events distant in space and time, the details and meaning of the behaviours of others, our innermost thoughts and desires, and so on."

There is a second specificity of language, linked not to its structure, but to its use in communication (sees Scott-Phillips 2015): the extent to which linguistic communication makes use of implicitly communicated content. This is acknowledged by both *Contextualists* (see, e.g., Carston 2002, Recanati 2004, Sperber and Wilson 1995, Wilson and Sperber 2012) and *Minimal Semantics* (see, e.g., Borg 2004, 2012, Stanley 2007).⁷ This, as we shall see in the

⁵ Note that this raises a subsidiary question: *Where does that infinity of different contents come from in the first place?*

⁶ The only exceptions are language and the honeybee dance. Because honeybees are *eusocial* insects that share a majority of genes, deception is not an issue in the honeybee dance. The situation is radically different in humans, in which communication is not limited to kin.

⁷ Roughly, Contextualists argue that the Gricean distinction (see Grice 1989) between *sentence meaning* (linguistically determined and corresponding to *what is said*) and *speaker meaning* (depending on the speaker's intentions and corresponding to *what is communicated*) should be abandoned. Both what is said and what is communicated fall under speaker meaning. By contrast,

following discussion, raises further difficulties for theories that see language as a communication system in the strong sense. Before we turn to that specific problem, I would like to explain why the two questions mentioned earlier could be answered more easily by an alternative theory and what, roughly, that alternative theory might be.

As said before, most theories of language evolution see language as a communication system in the strong sense that it evolved for communication. Yet, another view is possible, according to which *language did not evolve primarily for communication, but for another purpose, and was then exapted for communication*. Let us say that views that take this alternative path see language as a *communication system in the weak sense*. Such views have a fairly easy answer to the aforementioned questions: the core combination of features that is characteristic of language did not evolve for communication (where they do not make much sense), but for another purpose. They were inherited by language as a communication system in the weak sense. Hence, they do not have to be accounted for in a communicative framework. So the next question is: What did language evolve *for*, if not for communication?

First of all, let me preempt a potential objection to the effect that the very fact that language is routinely used in communication is a strong cue that it evolved *for* communication. In fact, present use may not be a good guide to primordial evolutionary function. A standard example is wings and feathers in birds. Nowadays, in most bird species, they are used for flight, which strongly suggests that flight was their primordial evolutionary function. However, wings and feathers in birds first evolved not for flight, but as a cooling system (see, e.g., Longrich et al. 2012), and were then *exapted* for flight.

So to return to what the primordial evolutionary function of language was, let me first point out that, while on a view that sees language as a communication system in the strong sense the two questions of why humans needed to communicate an infinity of different contents and of where that infinity of different contents come from are dissociated, this is not necessarily the case on the alternative view that language is a communication system in the weak sense. There the two questions can be merged, as we shall now see. The reason why language as a communication system allows us to communicate an infinity of different contents is that it has the core combination of discrete infinity, semanticity, and decoupling. This is the *basis* on which the primordial evolutionary function of language can be determined. What it suggests is that the reason for which humans can think of an infinity of different contents to communicate is that thought shares the same structural combination of features:

Minimal semanticists insist on the distinction (though they disagree on the characterization of what is said). Both sides acknowledge, however, the widespread role of implicit communication in what is communicated (for an enlightening discussion, see Borg 2012, Chapter 1).

discrete infinity, semanticity, and decoupling (as was recently argued by Fodor and Pylyshyn 2015, following Fodor 1975, 2008). And this combination of features that is such an embarrassment on an account according to which language evolved *for communication* makes perfect sense on an account according to which language evolved *for thought*. It explains where the infinity of contents that humans can communicate comes from. Additionally, decoupling is no problem in a system for thought, as the issue of deception does not arise as such. And the combination of discrete infinity, semanticity, and decoupling makes for a powerful system of thought such as that which humans enjoy (see Section 9.3 for a quick discussion), and that great apes arguably do not, or at least not to the same extent.

Let me now turn to the reason why implicit communication is a problem for views according to which language is a communication system in the strong sense. On such a view, communication systems are sets of signals specific to a species. Signals themselves are defined as *occurrent behaviors that have evolved in tandem with responses on the one hand and information on the other hand* (see Maynard Smith and Harper 2003). This is compatible with either a biological or a cultural evolutionary process.⁸ In most if not all animal communication systems, signals are more or less uniform across the species and are phylogenetically inherited (very little learning, if any, is necessary). In the human species, by contrast, signals *are not uniform across the species* (there are around 6,000 extant languages) and are subject to learning hence arise through ontogenetic learning. In other words, while most if not all animal communication systems are the product of biological evolution, human language is the product of cultural evolution.⁹ Basically, this means that the pairings between signals and information¹⁰ will have to be thought of along constructionist lines (see Goldberg 2006, Tomasello 2003), where signals are sentences or constructions (but not utterances), and where the information is to be understood as *what is communicated*. This is precisely because language is seen as a tool

⁸ Biological evolution takes place in geological time (hence it is slow), is due to natural selection or genetic drift, and either is neutral to or enhances fitness. By contrast, cultural evolution takes place in historical time (hence it is fast), is due to artificial selection or serendipity, and can lead to a “ratchet” effect (see Tomasello 2000). Notably, however, cultural evolution does not always lead to fitness (see Edgerton 1992). Note that, on the present dual account, the evolutionary process is *both* biological and cultural (see here, Janssen and Dediu).

⁹ Note that the question of the evolution of language is thereby reduced to the question of the evolution of E-languages (on the Chomskyan distinction between I-language and E-language, see Chomsky 1986). On the alternative view proposed earlier, according to which language is a communication system in the weak sense, that originally evolved for thought, by contrast, the question of language evolution is primordially that of the basis for I-language.

¹⁰ I will leave aside the pairings between signals and responses here, though for a discussion, see Reboul 2015.

for communication. What is relevant, hence, is what is communicated, i.e., on all accounts, speaker meaning. But, as mentioned earlier, speaker meaning is dependent on speaker's intentions, and thus unstable. Given that pairings are supposedly achieved by the observation of repeated correlations between signals and information, it seems unlikely that speaker meaning is stable enough to support the correlations in question.

Thus, the view that language is a communication system in the strong sense seems hardly tenable. Note, however, that adopting the view that it is a communication system in only a weak sense does not exempt us from giving an account of why it was exapted for communication. And, given that communication is the epitome of a social activity, this means giving a social account of that part of the process. Here one might want to amend one of the extant social scenarios (by dismissing the part according to which language is a communication system in the strong sense) and adopt it for the second step of the evolution of language as a communication system in the weak sense, i.e., for the social account of language exaptation for communication.

9.3 What is the Proper Social Account for the Exaptation of Language for Communication?

9.3.1 Introduction

Let me first point out that, beyond the specific problem of language evolution, the very evolution of communication, as long as *communication* is understood as information transfer (as is largely the case: see, e.g., Fitch 2010; Hauser 1996; Maynard Smith and Harper 2003; Oller and Griebel 2004, 2008), is a deep mystery. The problem, first pointed out by Krebs and Dawkins (1984), is that information is a precious commodity, so why share it? Sharing information, on the face of it, is *altruistic* in the sense that it is detrimental to the sender and beneficial for the recipient. So communicating, understood as sharing information, does not seem to make any evolutionary sense.

Krebs and Dawkins's (1984) solution was that communicating was not so much about sharing information as about inducing in the recipient a behaviour that is beneficial to the sender.¹¹ In their terms, communication was about manipulating the audience, and information transfer was largely incidental to it. This is why accounts of communication see it not only as information transfer but also as inducing a response in the audience and why definitions of the signal (the communicative unit) link it with *both* information and response (see

¹¹ For a development of this idea, setting resolutely aside the notion that animal communication is about sharing information, see Owren et al. (2010).

Maynard Smith and Harper 2003 and Section 9.2). As a communication system, language cannot entirely escape from such considerations.

However, as mentioned in Section 9.2, I will defend here a view according to which language is a communication system only in a weak sense, which entails a dual or two-step account, the second step having to do with the social exaptation of language for communication. As we saw, this means that structural properties of language (i.e., the core combination of discrete infinity, semanticity, and decoupling) do not have to be addressed at the second step of the account, as they are accounted for on the first step (see Section 9.2 for a short presentation of what such an account might be). It does not, however, mean that the problem raised by Krebs and Dawkins (1984) is innocuous at this second, social stage of the exaptation of language for communication. If anything, given that human responses seem too fickle for any strong correlations to linguistic signals, and given that language obviously and centrally conveys information, this problem is especially acute for language.

As we shall now see, most, if not all, social accounts of language evolution have, however, ignored the problem. Rather, they try to account for the *uniqueness* of language, which suggests that something is radically different between humans and other primates, including great apes. On a social account of language evolution, the change has to be in social circumstances, either in the prosocial attitudes of the human species relative to other primates or in the social organization of human groups.

9.3.2 *A Short Overview of Extant Social Scenarios for Language Evolution*

Some scenarios have homed in on the differences in life histories: humans, because of bipedalism and its consequences on anatomy (and notably on female anatomy), are born fairly immature (in any other primate species, they would be premature) and undergo a long period of altriciality during which they are entirely dependent on parental support for survival. The consequence is the formation of the human parental pair, which is an oddity among primates (where it is found only in a few species of New World monkeys) and unique among apes. This has led to two types of scenarios: the *motherese* scenario, where language developed to ensure long-distance contact when the mother has to put the baby down while foraging (see Falk 2009); and the parental pair bonding forced by female rituals devised to avoid male philandering and abandonment (the *bonding/ritual* scenario; see Power 1998). While both scenarios eschew more or less satisfactorily the problem raised earlier of the apparent altruism of communication (because, arguably, at least in the motherese scenario, the communicators' interests coincide, making deception unlikely), they seem far off the mark given the spread and openness of human communication. This is all

the more true given that humans still have an ancillary system of communication (presumably inherited from the last common ancestor with chimpanzees), made of inarticulate noises (such as laughter, grunts, shouts, etc.) and facial expressions indicating emotions. It is not clear why this would not have been sufficient for the purposes described earlier, given that all primates seem to have a range of vocal signals dedicated to mother-infant communication and that, in the one species where females are dominant (bonobos), they manage to keep males (see Furuiichi and Thompson 2008) where they want them to be without language. So let me now turn to the two most prominent social scenarios nowadays, those proposed by Dunbar (1996, 1998) and by Tomasello (2009, 2010). Tomasello's view is squarely on the side of a difference in the social make-up of humans relative to other primates (notably chimpanzees), while Dunbar argues for a change of social organization due to an increase in group size in modern humans.

Let me begin with Tomasello's (2009, 2010) story, as it is the simplest. Tomasello sees language evolution as stemming from a deep change in prosocial attitudes in humans. While apes (and notably chimpanzees) are competitive, humans are cooperative.¹² Why humans and only humans, should have done that switch from competition to cooperation is not entirely clear.

Tomasello suggests that this is due to greater cooperation not only in hunting (and here he refers to the *Stag Hunt* model; see Skyrms 2004) but also in gathering and describes a general sharing of the product of both activities. This is problematic, given the empirical evidence both relative to those chimpanzee societies where hunting is indeed cooperative (as it is in the chimpanzees of the Tai forest; see Boesch 1994a, 1994b, 2002, 2005; Boesch and Boesch 1989; Boesch & Boesch-Achermann 1991) and to human hunter-gatherer societies. In both of these cases, the product of the hunt is shared, but the product of gathering is not (see, e.g., Boehm 1999, for a general analysis of the social organization in such societies of hunter-gatherers). And clearly, while gathering in humans (though not in chimpanzees) may be a communal activity, it is not thereby cooperative, as no cooperation is needed. So there seems to be something of a gap in Tomasello's theory.

A more pressing worry is how far it is true that humans are altruistically cooperative. The main argument is that humans will go out of their way to help not only kin and friends but also strangers that they have just met and have very little chance of meeting again. This, or so it is claimed, excludes reciprocal altruism (which is a form of delayed mutualism, where both parties benefit in the end; see Trivers 2002). Thus, the only remaining explanation would

¹² And although Tomasello does not give an explicit definition of the term, it seems clear from what he says that he sees human cooperation as *altruistic*, i.e., detrimental to the agent and beneficial to the recipient.

be pure altruism, where the agent loses and the recipient gains. The problem is that while this may be true in advanced Western societies (where, incidentally, one might argue for a generalized reciprocal altruism, distributed on the whole group rather than operating on a one-to-one basis, and thus for a mutualistic rather than altruistic explanation), it seems a travesty of what actually is the case in less-developed societies, and notably in hunter-gatherer societies, where violence is endemic between and inside groups (see Keeley 1996, and in a more anecdotal vein, Chagnon 2013 and Diamond 2012). Meeting either with someone of a different group or a stranger is more likely to lead to violence than to helping, just as it is in chimpanzee societies. And it seems hard to deny that the “us against them” instinct in humans is still running strong (see Hardin 1995). Additionally, it may be doubted that humans are really so altruistic, given human history, including recent history even in developed countries (see, e.g., Kershaw 2008; Lifton 2011).

A second worry is whether it is true that humans are all that different from other apes and whether the slogan *Nasty apes, nice humans* really applies. Here, it is instructive to look at the following quotation by Maestripieri (2012, loc. 218):

We may think we have outgrown the conditions that govern the lives of other primates. We no longer live in the jungle and swing between trees; instead, our homes are in or around large cities, and we drive cars, wear clothes, spend years in formal education, and communicate electronically. Yet technology and clothes cannot disguise the inheritance of our primate past. They have simply changed the arena in which we act out age-old rituals, making the games that human primates play more arbitrary perhaps, but no less powerful.

Basically, what Maestripieri is saying is that the main difference between humans and other primates does not lie in their social abilities. Rather, it lies in the cognitive abilities that have led to humans driving cars, living in skyscrapers, communicating electronically, etc., while other primates are still living in forests.¹³ But regarding human societies, they still function, *mutatis mutandis*, along primate lines (for a fairly similar opinion, see De Waal 1998).

While Maestripieri is clearly on the Machiavellian side of primate sociality (see also Maestripieri 2007), other primatologists have argued that chimpanzees are able of altruistic behavior (e.g., males adopting orphans in Tai chimpanzees; see Boesch et al. 2010) and are also adept at reconciliation after in-group fights (see De Waal 1989; Wittig & Boesch 2005). Finally, by focusing

¹³ Note that this supports the view of language evolution proposed in Section 9.2. If language is, as is generally recognized, species specific and if the main difference between humans and other primates (including apes) lies in higher cognition rather than in social practices, it makes sense to suspect that there may be a deep link between higher cognition and language (rather than between social practices and language).

on chimpanzees and ignoring the other sister species to humans, i.e., bonobos, Tomasello may have warped the debate (see De Waal 2013). So, quite apart from the fact that there seems to be a long cry from the general benevolence that Tomasello sees as the prerogative of humans to language, the discrepancy that he advocates between humans and other primates, notably the great apes, may be more of a myth than a reality.

Dunbar's (1996, 1998) account is much richer in many ways than is Tomasello's, but is not entirely convincing either, as we shall now see. What is common between the two accounts is the idea that human intelligence is due to social pressures (what can be called the *Social Intelligence* hypothesis). According to Dunbar, language first evolved as a solution to the problem of securing social cohesion in groups of modern humans. In primates, group cohesion is usually secured through grooming. However, grooming is only possible for groups whose size is 120 members or less. This is because grooming is a costly activity and because the time one individual spends grooming other group members is directly correlated to group size: the bigger the group, the more time any individual will have to spend in grooming. Clearly, grooming interferes with other vital activities such as foraging for food and mating. Thus, there is a limit to the time any individual can devote to grooming, which is set at 30% of its time (see Aiello and Dunbar 1993). This is the limit reached in groups of 120 members. The consequence is that, in any bigger groups, grooming is not a practical solution to the problem of securing group cohesion. Hence, early modern human groups of 150 members had to find a substitute.

Dunbar's suggestion is that language was that substitute, allowing individuals to vocally "groom" as much as three other group members simultaneously, while leaving the hands free for other purposes. An obvious difficulty is that such vocal "grooming" could have been done by communal (and contentless) chanting (see, e.g., Mithen 2005, for such an account), and certainly does not need the core combination of discrete infinity, semanticity, and decoupling that characterizes language. In response to this obvious difficulty, Dunbar has added a Machiavellian twist to his story. Groups, becoming bigger, offer more opportunities for cheating, and this has led language to evolve and allow group members to exchange information about third parties and report their misdeeds (i.e., to *gossip*), protecting the group from cheaters.

There are problems with both parts of Dunbar's story, i.e., the language as grooming part and the gossip part, and here I follow the excellent discussion of Dunbar's view in Power (1998). A first objection has to do with language as a sort of ersatz grooming. Power points out that the main reason why grooming is efficient for group cohesion is that *it is costly*, thus reliably signaling the groomer's commitment to the groomee. By contrast, the vocal "grooming" provided by language is cheap and would be a very unreliable way of signaling commitment to the recipients. Regarding the idea that language evolved to

its full form (with the core combination of discrete infinity, semanticity, and decoupling) for gossip, it is subject to a not dissimilar objection: gossip, which is of necessity decoupled as it concerns absent third parties, far from being a tool against cheating, seems a tool made for cheaters. Nothing can prevent a cheater to denounce innocent third parties and escape with her own reputation intact.¹⁴

There is another problem with Dunbar's account in addition to those Power (1998) discussed and it has to do with the number of members in modern human groups. On the face of it, there seems to be no way to know exactly how many members groups of modern humans would have had around 150,000 to 200,000 years ago (when the species emerged). So how does Dunbar come up with his number of 150 per group? Dunbar's theory of language evolution is built on the foundation of his theory of brain increase in primates. According to him, brain size in primates is directly correlated with group size: thus, any increase in group size (due to changes in ecological conditions, notably to an increase in predation pressure) is only possible if it is accompanied by an increase in brain size (see Dunbar 1992).¹⁵ Given this strong correlation between group size and brain size in primates, Dunbar extrapolated group size for past species or for extant species in their distant past (i.e., *Homo sapiens*) on the basis of their brain sizes: thus, *Homo erectus*, at 120 members per group, would still have been able to make do with grooming, while early modern humans with 150 members per group would have had to resort to language.

There is, however, a major problem with that theory. A meta-analysis by Reader and Laland (2002) showed that enhanced brain size in primates is correlated not so much with group size as with cognitive abilities manifesting themselves in social learning, behavioral innovation, and tool use. This led Seyfarth and Cheney (2002, 4441) to the following conclusion: "Natural selection may, therefore, have favoured an increase in brain size because of benefits derived from innovation or social learning *that are independent of a species' typical group size*" (emphasis added). As mentioned earlier, group size is crucial to Dunbar's theory of language evolution, and the correlation between group size and brain size is crucial to Dunbar's estimate of group size in early modern humans. The whole theory rests on it. What can be concluded from Reader and Laland's meta-analysis is that this foundation is at best shaky and at worst nonexistent, shedding deep doubt on the whole enterprise.

So the extant social theories of language evolution do not seem to be entirely convincing. Nevertheless, even on a dual, two-step account of language

¹⁴ This agrees with the discussion in Section 9.2 and with the problem decoupling raises for accounts that see language as a communication system in the strong sense.

¹⁵ Dunbar (1992) supported this theory by plotting the average group size in extant primate species against the average brain size of those species.

evolution (such as the one proposed in Section 9.2), we do need an account of the second step, in which language was exapted for communication. And given that communication is the epitome of a social process, it seems commonsensical that such an account must be in part social.

9.3.3 Which Sort of Social Account Do We Need?

A first and not unimportant thing to note is that the notion of a major change in prosocial attitude does not seem quite right. Apart from the fact indicated in Section 9.3.2 that humans and great apes (if not primates in general) do not seem as socially different as would be the case if Tomasello were right, it is not clear why hypothetical human benevolence would lead to complex language. In addition, while it is true that it would somewhat lessen the evolutionary burden relative to (altruistic) communication (on this view, we would altruistically communicate because we *are* altruistic), it raises the even more difficult problem of how altruism would have evolved in the first place. On the face of it, how altruism can evolve is an unsolved problem in the theory of evolution, though it may be argued that it does not arise in the first place because altruism just does not exist in the form Tomasello proposes. Arguably, in all the evolutionary theories of cooperation, cooperation is not altruistic, but self-serving in one way or another.¹⁶ So altruism does not seem a viable proposal.

Yet, as noted by Fitch (2011, 142), “humans . . . have an irrepressible habit of sharing their thoughts to others.” Fitch proposes to call this very human propensity by the composite German word *Mitteilungsbedürfnis*. This raises the very legitimate question of what, if not altruism, is the root of human *Mitteilungsbedürfnis*? It is this question to which we will now turn.

It is important to frame the debate over the social step in human learning in a fairly precise way (something that is lacking in most social scenarios, which seem to be content to use terms in a vernacular and often imprecise way). So let me begin by distinguishing between types of *social activities* on the one hand and types of *social attitudes* on the other hand. There are three types of social activities I am interested in: *collaboration*, *cooperation*, and *manipulation*. And I distinguish between four types of social attitudes: *mutualism*, *altruism*, *exploitation*, and *free-riding/selfishness*. All of these can be defined in “economic” terms that are compatible with biological/evolutionary definitions.

¹⁶ In *kin selection* theory (see Hamilton 1964a, 1964b), organisms will help their kin because kin share at least part of their genes, and thus helping kin means increasing one’s own inclusive fitness (i.e., the chance of passing on their genes). In *reciprocal altruism* theory (see Trivers 2002), agents are helpful to the extent that they will be helped in return. Neither of these views has anything to say about the kind of “pure” altruism proposed by Tomasello.

So let me begin with social activities:

- *Collaboration*: Occurs when an activity is performed by several individuals, where the behavior of each participant is coordinated with the activity of the other participants, and where the participants have a common goal;
- *Cooperation*: Occurs when an agent does something beneficial for the recipient;
- *Manipulation*: Occurs when an action by an agent toward a recipient leads the recipient to perform a behavior that the recipient might otherwise not have performed and that is beneficial for the agent.

Social attitudes are manifested through social activities, but are supposed to correspond to the tendencies an agent has for involving itself into this or that type of social activity:

- *Mutualism*: The activity is directly beneficial to all participants.
- *Altruism*: The activity is beneficial to the recipient but detrimental to the agent.
- *Exploitation*: The activity is beneficial to the agent and neutral for the recipient.
- *Free-riding/Selfishness*: The activity is beneficial to the agent and detrimental to the recipient.

Note that, in and off themselves, these definitions are *functional* in the sense that they do not presuppose any awareness in the agent. In other words, they can apply to any animal activity, not only to human activity. In human activity, however, one would expect both intentionality and awareness, in at least a majority of cases. Given these definitions, one can say that collaboration will be mutualistic, while cooperation can be either mutualistic or altruistic. Manipulation will be exploitative, selfish, or mutualistic.

As mentioned earlier, quite a few social accounts of language evolution assume that linguistic communication is cooperative and based on a deep altruism in humans. In other words, language has evolved in humans (and only in humans) because humans have shifted from the competitive, and hence exploitative or selfish, social attitudes that are characteristic of primates (and, notably, of chimpanzees) to an altruistic stance, which manifests itself in linguistic cooperation (see, e.g., Tomasello 2009, 2010, 2014 for an explicit account along such lines). Setting aside for the sake of the argument the objections that were already given earlier in the chapter to such a shift in humans, what reason do we have to believe that language is altruistically cooperative?

The view seems to derive from a very specific, and, as we will see, misguided, interpretation of Grice's *Logic of conversation* (see Grice 1989). As is well known, Grice's Logic of conversation aimed at accounting for the derivation of

a particular type of implicit content, i.e., *implicatures*. A standard example of implicature is:

- (1) Anne lives somewhere in Burgundy.
- (2) The speaker does not know exactly where Anne lives.

While (1) explicitly communicates that Anne lives in Burgundy (this is its *sentence meaning* in Gricean terminology), it implicitly communicates the content in (2), which is an implicature (and its *speaker meaning* in Gricean terminology). According to Grice, while sentence meaning is recovered through purely linguistic processes of interpretation (and is indeed the result of semantic compositionality), speaker meaning is recovered through pragmatic inferential processes. The central premise in these inferential processes is that the speaker has complied with a general principle that constrains *rational* communication among humans, the *Cooperative Principle*¹⁷:

Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged.

There is a fairly general consensus nowadays that, indeed, Grice was right, and that linguistic communication cannot only be a matter of encoding (in a linguistic code, i.e., an E-language) and decoding a message. Pragmatic inferential processes have to come in and they play an important role in utterance interpretation. Given this semantic underdetermination, it seems clear that some sort of cooperation has to occur: the hearer can only recover speaker meaning if the speaker has cooperated in the sense of making it possible for the hearer to recover her meaning, by tailoring her utterance to (what she believes to be) her hearer's abilities (i.e., his knowledge and beliefs). This can be seen not only in implicatures but also even in the choice of referential expressions:

- (3) *Magali's father* will come tomorrow.
- (4) *Jean's brother* will come tomorrow.
- (5) *Peter's son* will come tomorrow.
- (6) *Marianne's ex-husband* will come tomorrow.
- (7) *The village butcher* will come tomorrow.

Suppose all these referential expressions (indicated by italics) refer to the same man. They will be used for different hearers, taking into account these hearers' respective knowledge of the man's family relations and professional status. So cooperation is indeed important for linguistic communication. But does that mean that the cooperation involved in linguistic communication is altruistic?

¹⁷ The Cooperative Principle is accompanied by maxims. I will not get into that level of detail.

As we saw from an earlier discussion, cooperation is not necessarily altruistic. It can also be mutualistic. So is human linguistic communication based on altruistic or on mutualistic cooperation?

Just as a quick reminder, if human communication is a mutualistic activity, both the speaker and the hearer benefit from it, while if it is an altruistic activity, only the hearer benefits, and the speaker loses, from it. In human communication, both the speaker and the hearer would seem to have the common goal of exchanging information. In other words, it is in the speaker's interest to tailor her utterance to facilitate the hearer's interpretive process and in the hearer's interest to recover the speaker meaning. Thus, both can be said to benefit. Hence, on the face of it, the cooperation involved is mutualistic rather than altruistic. This, however, leaves us with the deeper question raised by Krebs and Dawkins (1984) (see Section 9.3.1) about why the speaker would want to share information in the first place. Or, in Fitch's (2011) words, where does human *Mitteilungsbedürfnis* come from? Isn't that where altruism gets into human linguistic communication?

Here, it is worth pointing out that indeed the question of cooperation in human linguistic communication and the question of human *Mitteilungsbedürfnis* are not only distinct in fact, but should also be distinguished in principle. In evolutionary theory, there is a standard distinction between *ultimate* and *proximate* explanations. The standard example is sex:

- Sexual intercourse is pleasurable and this *proximately* explains why humans have sex.
- Sexual intercourse is (or was until recently) the only means of reproduction and this *ultimately* explains why humans have sex.

Note that the ultimate explanation also grounds the proximate explanation. Regarding human linguistic communication, mutualistic cooperation explains *how* it can take place. It is a proximate explanation. But it does not explain *why* it takes place. For that we need an ultimate explanation.¹⁸

So what is the ultimate explanation for linguistic communication in humans, the explanation for why language was exapted for communication after having evolved for thought? Note that this question is made even more pressing by the fact that, arguably, humans still have a primate communication system, on a par with that of chimpanzees, made of inarticulate vocalizations, postures, laughter, facial expressions, etc. (see Burling 2005). So why would they have needed an additional and much more powerful system of communication? An obvious answer is that they had much more to communicate about than do other primates, precisely because they had a much more powerful system of thought.

¹⁸ Arguably, Tomasello's mistake in offering his altruistic cooperation view is in conflating these two different levels of explanation.

This will not quite do, however. Animal communication is on the whole very limited. No animal communication system seems to go over a limit of 30 to 40 signals, while, despite the difference in magnitude between animal and human conceptual apparatuses, there does not seem to be such a low limit on the number of concepts animals have. The conclusion, which seems fairly consensual (see, e.g., Fitch 2010; Hurford 2007), is that what animals communicate does not do justice to their cognitive abilities.¹⁹ In other words, animals do not share human *Mitteilungsbedürfnis*. And this leads us to the conclusion that there is no automatic link between richness of thought and richness of communication.

So let us go back one step and examine more carefully the point on which the use of language in communication seems utterly different from what happens in animal communication, setting aside the structural differences between language and other animal communication systems (i.e., discrete infinity, semanticity, and decoupling), which are not in need of an explanation at that level, as they are already accounted for. This feature, unique to linguistic communication, is implicit communication. It is unique in the sense that, as far as we know, it is universal in languages (see von Finkel and Matthewson 2008), it is not found in any other animal communication system (see Hauser 1996; Fitch 2010), and it does not seem to be part of Universal Grammar. As mentioned earlier, linguistic communication is rife with implicit communication. Indeed, if both Contextualists and Minimal Semanticists are right, even what is said (non vacuous truth-conditional, propositional meaning) is context dependent to a degree. But what I am interested in here, beyond disambiguation and saturation processes, are the two main forms of implicit communication that were identified in the second half of the 20th century by, respectively, Strawson (1950) and Grice (1975/1989), i.e., *presuppositions* and *implicatures*.

Beginning with presuppositions, Strawson (1950) noted that some utterances communicate a content that is not truth-conditional in the sense that it is not affected by modifications that impact the truth-conditions of the proposition communicated by the utterance, such as negation or interrogation. Thus, the utterance (8) explicitly conveys the proposition in (9). It also conveys the content in (10). When affected by negation, as in (11), the proposition it explicitly conveys is modified as in (12). And when affected by interrogation as in (13), its truth-value is suspended. Nevertheless, (8), (11), and (13) still convey (10):

- (8) John has stopped drinking.
- (9) John does not drink. [*main content*]
- (10) John used to drink. [*presupposition*]
- (11) John has not stopped drinking.

¹⁹ Note that, despite claims to the contrary, this does *not* entail that animal cognition is on a par with human cognition.

- (12) It is not the case that John does not drink [*main content*]
- (13) Has John stopped drinking?

Strawson called this additional non-truth-conditional content a *presupposition*.

We already know from an earlier discussion what an implicature is. I just reproduce the example in (1)–(2):

- (1) Anne lives somewhere in Burgundy.
- (2) The speaker does not know exactly where Anne lives.

Arguably, the central question regarding implicit communication in this restricted sense (i.e., limited to presuppositions and implicatures) is why we have it at all. One could argue that linguistic form and/or lexical meaning, though not truth-conditional, mandate presuppositions. It seems rather more difficult to say the same about implicatures: there is nothing linguistic in (1) to mandate the implicature in (2), and indeed Grice proposed that (2) is recovered through a reasoning based on the fact that if the speaker does not say where exactly Anne lives, it is probably because she does not know.

Regarding implicatures, there may be two different answers (baring the Universal Grammar account):

- Implicatures are an automatic result of general principles that govern the use of language in communication (e.g., *Minimax* principles).
- Implicit communication is used for social reasons, as communicative strategies (and this second explanation covers presuppositions as well as implicatures).

Beginning with the first answer, Minimax principles are basically economic principles to the effect that costs should be *minimized* while benefits should be *maximized*. In contemporary linguistics, two sorts of accounts have appealed to minimax to explain implicatures: the *neo-Gricean* approaches (e.g., Horn 2004; Levinson 2000) and the *post-Gricean* approaches (basically Relevance Theory; see Sperber and Wilson 1995). Their use of Minimax principles is fairly different, however.

Horn's Minimax is both speaker- and hearer-oriented, and Horn insists on the importance of *form* rather than content in the derivation of implicatures, notably in his discussion of *categorical sentences*:

- A: All/Every F is G.
- E: No F is G.
- I: Some F is/are G.
- O: Not every F is G/Some F is not G.

As is well known, A-sentences imply (in the sense of logical or material implication) I-sentences, while E-sentences imply O-sentences. But categorical sentences are also related by implicature relations: O-sentences implicate the negation of E-sentences, while I-sentences implicate the negation of A-sentences. Horn explains these implicature relations by the fact that the “basic” forms (A/E) are *briefer* (rather than merely more informative) than the I/O forms. Hence, using the I/O forms implicate against the A/E interpretations.

There are two main problems with Horn’s views. The first is that it can only account for a subset of implicatures: those that are linked to *lexical scales*, in which lexical items are constrained by the same implication and implicature relations as are the quantifiers in categorical sentences. Such implicatures are called *scalar implicatures*. But, clearly, Horn cannot account for the other implicatures, illustrated by (1) and (2) mentioned earlier. The second problem is that Horn’s account supposes that linguistic production is costly while pragmatic inference (necessary to derive the implicature) is cheap. This view, however, is contradicted by experimental evidence that has shown that for a sentence such as (14), the semantic interpretation in (15) is more quickly accessed than the pragmatic interpretation (the implicature) in (16):

- (14) Some elephants are mammals.
- (15) Some and maybe all elephants are mammals. [*semantic interpretation*]
- (16) Some, but not all elephants are mammals. [*pragmatic interpretation*]

What is more, in Horn’s account, where linguistic form is the main factor involved, one would expect the pragmatic interpretation to be automatically derived. This, however, is not what is found: The semantic interpretation is given in 40% of cases, which means that pragmatic interpretations are not at ceiling (if it were automatic indeed, one would expect it to be given at rates of 90–100%).²⁰ So Horn’s account is limited to a subset of implicatures and fails to account for even that subset.

Relevance Theory (the post-Gricean alternative) is also based on a Minimax principle, the principle of Relevance that basically says that an utterance is relevant to the extent that its interpretive costs are low and its cognitive effects important. Note that, by contrast with Horn’s Minimax, the Relevance-theoretical Minimax is hearer-based and has nothing to say about the speaker. The central idea is that any utterance conveys the presumption that its cognitive effects will balance its interpretive costs. Relative to implicature and given that Relevance Theory sees pragmatic inferences as costly, this means that extra-cognitive effects have to be present.

²⁰ For the experimental work, see, e.g., Noveck 2001, Bott and Noveck 2004, Bott et al. 2012.

Let us look at example (17):

- (17) A: Do you want some wine?

B: No, thank you.

B': I don't drink alcohol.

The answer in B is explicit, while the answer in B' is implicit. The implicature is that B does not want any wine. However, recuperating that content entails additional interpretive efforts on A's part, as she has to do a pragmatic inference to the effect that wine is an alcoholic beverage and that if B is a teetotaller, he will not want to drink wine. So there has to be additional cognitive effects to offset the additional interpretive costs and Sperber and Wilson note that B' also communicates the following set of assumptions (which are not communicated by B's explicit answer):

- (18) B doesn't drink beer.

B doesn't drink vodka.

B doesn't drink whisky.

Etc.

In other words, Sperber and Wilson's argument is based on a comparison between utterances with *both* different interpretive costs *and* different cognitive effects. However, this seems less than conclusive. The correct comparison would be between two utterances with similar cognitive effects but different costs, for instance an explicit answer such as (19) and an implicit answer such as (20):

- (19) B: No thanks I don't drink alcohol.

- (20) B': Thank you, I don't drink alcohol.

There is no reason to think that the hearer of (19) would not recover the assumptions in (18). In addition, recent experimental work on scalar implicatures, comparing time reactions for the pragmatic interpretation of utterances using *some* and for the semantic interpretation of explicit utterances using *only some*, has shown that explicit communication is faster (less costly) than pragmatic interpretation, keeping cognitive effects constant (see Bott et al. 2012). Note that this does not mean that optimising relevance by looking for extra cognitive effects is not the right explanation for *how* implicatures are accessed. Rather it means that it is *not* a valid explanation for *why* we have implicatures in the first place. One might want to argue (as Sperber and Wilson actually do) that utterances with implicatures are produced because there was no other way of conveying the implicated content.²¹ However, this is clearly not the case for all

²¹ Arguably this is the case for live metaphors (see Reboul 2014).

implicatures, as shown by examples (19) and (20) mentioned earlier, and it certainly is not the case for scalar implicatures.

Thus, both types of Minimax accounts give, at best, a proximal explanation of implicatures, but not an ultimate explanation. My suggestion is that an ultimate explanation not only for implicatures but for all implicit communication (i.e., implicatures and presuppositions) would take us a long way toward an ultimate explanation of the existence of human linguistic communication and of human *Mitteilungsbedürfnis*. So what is the ultimate explanation for implicit communication?

Here, we turn to the idea that the reason for which we have implicit communication is social and that the use of implicit communication in linguistic communication reflects deep speaker's strategies. While linguistic communication is cooperative in a mutualistic way *in order to secure communicative success*, this does not mean that cooperation, mutualistic or otherwise, is what motivates the use of implicit communication in the first place. Let me first define communicative success as the recovery by the hearer of a message that is sufficiently similar to the speaker's thought to ensure that the effect intended can occur. The notion of an intended effect comes from Grice's definition of non-natural meaning (or meaning_{NN}) (see Grice 1989, 219):

“A meant_{NN} something by *x*” is roughly equivalent to “A intended the utterance of *x* to produce some effect in an audience by means of the recognition of that intention.”

This definition implies a double intention:

- The speaker's *primary* intention to produce a given effect in her audience;
- The speaker's *secondary* intention to produce that effect via the audience's recognition of her (the speaker's) primary intention.²²

Let me now turn to what happens when the speaker is lying. In such a case, the effect she intends to produce in her hearer is presumably exactly the same as the effect she would intend to produce if she were telling the truth: i.e., for an assertion, she presumably intends to produce in her hearer a belief to the effect that the propositional content of her utterance is true (or a belief whose content is that proposition). So in both lying and sincere assertion, the speaker's primary and secondary intentions are identical.²³

²² Sperber and Wilson (1995) replace Grice's primary and secondary intentions by *informative* and *communicative* intentions. The main thing to note is that dual-intention accounts, such as Grice's and Sperber and Wilson's, postulate such dual intentions to explain how speaker meaning is recovered. Arguably, accounts that postulate a single intention make pragmatic inference production a mystery (see Reboul and Moeschler 1998).

²³ Note that this is as it should be: arguably, lying can only be successful if the lie is not detected as such. Clearly, if the hearer had to detect in the speaker a primary intention to mislead him (the hearer), lies would always fail.

Yet, obviously, lying is an intentional act (it implies, pace Carson 2010, an intention to deceive), so there has to be a possibility for further intentions beyond the primary and secondary intentions in an act of communication. Let us call the Gricean primary and secondary intentions (or, equally, Sperber and Wilson's [1995] informative and communicative intentions) *proximal* intentions. Proximal intentions have to be recognized if communication is to be successful. The further intentions that are involved in, e.g., lying are *distal* intentions and do not have to be recognized to be fulfilled (indeed, in some cases at least, they *must not* be recognized if they are to be fulfilled).

So in addition to the dual proximal intentions, communicators can also have distal intentions. And, while the proximal intentions are cooperative in a Gricean and mutualistic sense, the distal intentions do not have to be cooperative in any sense. Remember that Krebs and Dawkins (1984) claimed that communication only makes sense if it is manipulative in the sense that it leads the audience to act in a way that is beneficial to the communicator. Arguably, the intention to manipulate is a distal intention that, more often than not, must not be recognized to be fulfilled. So the view I defend here is that linguistic communication is often manipulative, and that implicit communication emerged to allow the speaker to hide her manipulative intentions and to deny having had such manipulative intentions. Before we turn back to examples, I would like to make clear that manipulation does not necessarily involve deception in the content communicated (i.e., it can and often will be true) and that it need not be detrimental to the hearer (it may be neutral or even beneficial to him).²⁴

So let us go back to examples, beginning with implicatures:

- (21) A: Do you know where Anne lives?
 B: Somewhere in Burgundy, I believe.

As mentioned earlier, the explicit content of B's answer is that Anne lives in Burgundy. The implicit content is that B does not know where exactly Anne lives. Note that B may know Anne's precise address, but not want A to write to her or visit her. *Importantly, by her utterance, B can hide her intention not to give that information to A and deny having had that intention.*

Presupposition leads to very much the same conclusion:

- (22) A: I have decided to appoint John as the manager of the new store.
 B: That's an excellent idea, especially now that John has stopped drinking.

²⁴ The only requirement on manipulation, according to the aforementioned definition, is that it should lead to a benefit for the communicator. Note that this rules out the difficulty of the negative impact of detrimental effects on the recipient and its consequence of blocking the evolution of communication (see Maynard Smith and Harper 2003, as well as Section 9.2).

The explicit content of B's answer is that A's choice is a good one and that John does not drink. The presupposition is that John used to drink, an information that could lead A to change her mind. *Note, however, that B can claim that she has no ill will toward John, and point out that she has, indeed, praised A's choice.*

In none of the aforementioned cases is there any lie *stricto sensu*. Again, in both (21) and (22), the manipulation need not be detrimental to the hearer. And, finally, the intention the speaker can deny having is not a proximal intention, but a distal intention. Note that this analysis of implicit communication agrees with Pinker's *Theory of the Strategic Speaker* (see Lee and Pinker 2010; Pinker 2007; Pinker et al. 2008).

But is that all there is to implicit communication? One can, in addition, claim that implicit communication also allows the speaker to bypass her hearer's *epistemic vigilance* (see Sperber et al. 2010). Epistemic vigilance is part of linguistic communication (on the hearer's side) for two reasons:

- The interests of the speaker and those of the hearer may not coincide and the speaker may be deceptive or manipulative.
- According to the *Argumentative Theory of Reasoning* (see Mercier 2009; Mercier and Sperber 2011), human reasoning is first and foremost dedicated to argumentation, i.e., to convincing one's hearer that one's opinion is better than his (the speaker's) own. This entails a form of coevolution between linguistic communication and human reasoning (as partly distinct from thought). It also implies that reasoning involves both the production of reasons and *the evaluation of reasons produced by others*.

This second component of human reasoning is already a tool for epistemic vigilance. Additionally, the Argumentative Theory of Reasoning explains an often-noted characteristic of human reasoning, the *egocentric bias*, i.e., a preference for one's beliefs over communicated beliefs, which ensures that credibility is kept in check. What the egocentric bias seems to show is that if the hearer detects that the speaker is trying to change his beliefs, the hearer will immediately increase his distrust. There are a few consequences of the egocentric bias that seem directly relevant to the usefulness of implicit communication:

- People will be less vigilant toward information *that is not presented as a reason for the speaker to change his opinion or decision*, i.e., that is not presented as communicated with a manipulative or argumentative distal intention.
- They will also tend to be less epistemically vigilant toward information *on the truth of which the speaker does not seem to commit herself*.

Arguably, these are features of implicit communication. As we have already seen, both presuppositions and implicatures allow the speaker to hide or deny her manipulative (or argumentative) intention. The second condition for

decreased epistemic vigilance is also fulfilled by implicit communication. In presupposition, the presupposed content is “backgrounded” – presented as already known and consensual and as not part of the *question under discussion* (for that notion, see Roberts 2004). Regarding implicatures, the lack of speaker’s commitment is due to *defeasibility* (a feature of implicatures already noted by Grice 1989). Implicatures can be contradicted in the very utterance that seems to communicate them, without the whole utterance being contradictory or weird:

- (23) Anne lives somewhere in Burgundy, in Cluny in fact.

This means that rather than being content *asserted* by the speaker,²⁵ implicatures are conclusions that the hearer reaches himself, his *own conclusions*, so to speak, and thus it should be content that is favored by the egocentric bias.

So it seems that implicit communication is a tool for more or less manipulative argumentation. If this is the case, given how widespread implicit communication is in human linguistic communication, and given that it is specific to it, this suggests that the ultimate explication for why language was exapted for communication in humans is that language evolved for argumentation in groups where both interests and opinions could differ. In conclusion, I would like to say a few words about the anthropological circumstances where this happened.

9.4 Conclusion

Clearly, we cannot know with any certainty when exactly language emerged in hominines. However, there are a few clues that language as a communicative system emerged in modern *Homo sapiens*, and maybe that it only emerged about 70,000 years ago just before modern humans left Africa, when signs of intellectual creativity, symbolic activities, and diversified tool construction and use appear. Basically, this means that language appeared as a communicative system in groups of hunter-gatherers. What we know of such contemporary groups is that they are strongly egalitarian, presumably for economic reasons, which probably also applied in the early groups in which language evolved as a communication system. This egalitarian ethos is not due to a change in prosocial attitudes, but rather because such groups function as *reverse hierarchies* (see Boehm 1999). In reverse hierarchies, the group as a whole acts to prevent dominance by a single individual or subgroup. Any upstart will be socially sanctioned by sarcasm, social ostracism, and, ultimately, collective murder. In such societies, the assembly of males takes decisions above family level collectively. In such circumstances, persuasion and modesty rather than coercion and

²⁵ Which explains why, while a speaker can *deceive* via an implicature, she cannot *lie* via an implicature.

assertiveness are the rule. These are obviously circumstances in which both the Argumentative Theory of Reasoning and the emergence of implicit communication as a linguistic tool to persuade others without falling foul of epistemic vigilance make perfect sense.

References

- Aiello, Leslie C. & Dunbar, Robin I. M. 1993. 'Neocortex size, group size and the evolution of language', *Current Anthropology* 34(2): 184–193.
- Boehm, Christopher 1999. *Hierarchy in the forest: The revolution of egalitarian behavior*. Cambridge, MA: The MIT Press.
- Boesch, Christophe 1994a. 'Chimpanzees-red colobus monkeys: a predator-prey system', *Animal Behavior* 47: 1135–1148.
- Boesch, Christophe 1994b. 'Cooperative hunting in wild chimpanzees', *Animal Behavior* 48: 653–667.
- Boesch, Christophe 2002. 'Cooperative hunting roles among Taï chimpanzees', *Human Nature* 13(1): 27–46.
- Boesch, Christophe 2005. 'Joint cooperative hunting among wild chimpanzees: taking natural observations seriously', *Behavioural and Brain Sciences* 28(5): 692–693.
- Boesch, Christophe and Boesch, Hedwige 1989. 'Hunting behavior of wild chimpanzees in the Taï national park', *American Journal of Physical Anthropology* 78: 547–573.
- Boesch, Christophe and Boesch-Achermann, Hedwige 1991. 'Dim forest, bright chimps', *Natural History* 1991: 50–56.
- Boesch, Christophe, Bolé, Camille, Eckhardt, Nadin, and Boesch, Hedwige 2010. 'Altruism in forest chimpanzees: the case of adoption', *PLOS One* 5(1): e8901.
- Borg, Emma 2004. *Minimal semantics*. Oxford: Oxford University Press.
- Borg, Emma 2012. *Pursuing meaning*. Oxford: Oxford University Press.
- Bott, Lewis, Bailey, Todd M., and Grodner, Daniel 2012. 'Distinguishing speed from accuracy in scalar implicatures', *Journal of Memory and Language* 66: 123–142.
- Bott, Lewis and Noveck, Ira 2004. 'Some utterances are underinformative: The onset and time course of scalar inferences', *Journal of Memory and Language* 51(3): 437–457.
- Burling, Robbins 2005. *The talking ape: How language evolved*. Oxford: Oxford University Press.
- Byrne, Richard W. and Whiten, Andrew (eds.) 1988. *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes and humans*. Oxford: Clarendon.
- Carson, Thomas 2010. *Lies and deception: Theory and practice*. Oxford: Oxford University Press.
- Carston, Robyn 2002. *Thoughts and utterances: The pragmatics of explicit communication*. Oxford: Blackwell Publishing.
- Chagnon, Napoleon A. 2013. *Noble savages: My life among two dangerous tribes – The Yanomamö and the anthropologists*. New York: Simon and Schuster.
- Chomsky, Noam 1986. *Knowledge of language: Its nature, origin and use*. New York: Praeger.

- De Waal, Frans 1989. *Peacemaking among primates*. Cambridge, MA/London: Harvard University Press.
- De Waal, Frans 1998. *Chimpanzee politics: Power and sex among chimpanzees*. Baltimore: John Hopkins (Revised edition).
- De Waal, Frans 2013. *The bonobo and the atheist: In search of humanism among primates*. New York: Norton and Co.
- Diamond, Jared 2012. *The world until yesterday: What can we learn from traditional societies?* London: Penguin Books.
- di Scullio, Anna Maria and Boeckx, Cedric (eds.) 2011. *The biolinguistic enterprise: New perspectives on the evolution and nature of the human language faculty*. New York: Oxford University Press.
- Dunbar, Robin I. M. 1992. 'Neocortex size as a constraint on group size in primates', *Journal of Human Evolution* 20: 469–493.
- Dunbar, Robin 1996. *Gossip, grooming and the evolution of language*. London: Faber and Faber.
- Dunbar, Robin 1998. 'Theory of mind and the evolution of language', in Hurford, Studdert-Kennedy and Knight (eds.), pp. 92–110.
- Edgerton, Robert B. 1992. *Sick societies: Challenging the myth of primitive harmony*. New York/Toronto: The Free Press.
- Falk, Dean 2009. *Finding our tongues: Mothers, infants and the origins of language*. New York: Basic Books.
- Fitch, Tecumseh 2010. *The evolution of language*. Cambridge/New York: Cambridge University Press.
- Fitch, W. Tecumseh 2011. "Deep homology" in the biology and evolution of language', in di Scullio and Boeckx (eds), pp. 135–166.
- Fodor, Jerry A. 1975. *The language of thought*. New York: Thomas Y. Crowell.
- Fodor, Jerry 2008. *LOT2: The language of thought revisited*. Oxford: Clarendon Press.
- Fodor, Jerry A. and Pylyshyn, Zenon W. 2015. *Minds without meanings: An essay on the content of concepts*. Cambridge, MA: The MIT Press.
- Furuichi, Takeshi and Thompson, Jo (eds) 2008. *The bonobos: Behavior, ecology and conservation*. New York: Springer.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- Greenberg, Joseph (ed.) 1963. *Universals of language*. Cambridge, MA: The MIT Press.
- Grice, H. Paul 1989. *Studies in the way of words*. Cambridge, MA: Harvard University Press.
- Hamilton, William D. 1964a. 'The genetical evolution of social behaviour. I', *Journal of Theoretical Biology* 7: 1–16.
- Hamilton, William D. 1964b 'The genetical evolution of social behaviour. II', *Journal of Theoretical Biology* 7: 17–52.
- Hardin, Russell 1995. *One for all: The logic of group conflict*. Princeton: Princeton University Press.
- Hauser, Marc 1996. *The evolution of communication*. Cambridge, MA: The MIT Press.
- Hockett, Charles F. 1963. 'The problem of universals in language', in Greenberg (ed.), pp. 1–29.
- Horn, Laurence R. 2004. Implicature. In Horn and Ward (eds.), pp. 3–28.

- Horn, Laurence R. and Ward, Gregory (eds.) 2004. *The handbook of pragmatics*. Oxford: Blackwell.
- Hurford, James R. 2007. *The origins of meaning: Language in the light of evolution*. Oxford: Oxford University Press.
- Hurford, James R., Studdert-Kennedy, Michael, and Knight, Christopher (eds.) 1998. *Approaches to the evolution of language: Social and cognitive bases*. Cambridge: Cambridge University Press.
- Janssen, Rick and Dediu, Dan 2016. Genetic biases affecting language: what do computer models and experimental approaches suggest? In Poibeau and Villavicencio (eds.) *Language, Cognition, and Computational Models*. Cambridge University Press.
- Keeley, Lawrence H. 1996. *War before civilization: The myth of the peaceful savage*. New York/Oxford: Oxford University Press.
- Kershaw, Ian 2008. *Hitler, the Germans and the final solution*. Jerusalem/New Haven, CT: Yad Vashem/Harvard University Press.
- Krebs, John R. and Davies, Nicholas B. (eds.) 1984. *Behavioural ecology: An evolutionary approach*. Sunderland, MA: Sinauer Associates.
- Krebs, John R. and Dawkins, Richard 1984. 'Animal signals: mind-reading and manipulation', in Krebs and Davies (eds.), pp. 380–402.
- Lee, James L. and Pinker, Steven 2010. 'Rationales for indirect speech: the theory of the strategic speaker', *Psychological Review* 117 (3): 785–807.
- Levinson, Stephen 2000. *Presumptive meanings: The theory of generalized conversational implicatures*. Cambridge, MA: The MIT Press.
- Lifton, Robert J. 2011. *Witness to an extreme century: A memoir*. New York: Free Press.
- Longrich, Nicholas R., Vinther, Jakob, Meng, Qingjin, Li, Quqngguo, and Russell, Anthony P. 2012. 'Primitive wing feather arrangement in *Archaeopteryx lithographica* and *Andriornis huxleyi*', *Current Biology* 22: 2262–2267.
- Maestripieri, Dario 2007. *Machiavellian intelligence: How Rhesus Macaques and humans have conquered the world*. Chicago/London: The University of Chicago Press.
- Maestripieri, Dario 2012. *Games primates play: An undercover investigation of the evolution and economics of human relationships*. New York: Basic Books.
- Maynard Smith, John and Harper, David 2003. *Animal signals*. Oxford: Oxford University Press.
- Mercier, Hugo 2009. 'La Théorie Argumentative du Raisonnement', unpublished Ph.D. thesis, E.H.E.S.S, Paris.
- Mercier, Hugo and Sperber, Dan 2011. 'Why do humans reason? Arguments for an Argumentative Theory', *Behavioral and Brain Sciences* 34(2): 57–111.
- Millikan, Ruth 1984. *Language, thought, and other categories*. Cambridge, MA: The MIT Press.
- Millikan, Ruth G. 2004. *Varieties of Meaning: The 2002 Jean Nicod Lectures*. Cambridge, MA: The MIT Press.
- Millikan, Ruth G. 2005. *Language: A biological model*. Oxford: Clarendon Press.
- Mithen, Steven 2005. *The singing Neanderthals: The origins of music, language, mind and body*. London: Orion Books.
- Noveck, Ira 2001. 'When children are more logical than adults: experimental investigations of scalar implicature', *Cognition* 78(2): 165–188.

- Oller, D. Kimbrough and Griebel, Ulrike (eds.) 2004. *Evolution of communication systems*. Cambridge, MA: The MIT Press.
- Oller, D. Kimbrough and Griebel, Ulrike (eds.) 2008. *Evolution of communicative flexibility: Complexity, creativity and adaptability in human and animal communication*. Cambridge, MA: The MIT Press.
- Owren, Michael J., Rendall, Drew, and Ryan, Michael J. 2010. ‘Redefining animal signaling: influence versus information in communication’, *Biology and Philosophy* 25: 755–780.
- Pinker, Steven 2007. ‘The evolutionary social psychology of off-record indirect speech acts’, *Intercultural Pragmatics* 4(4): 437–461.
- Pinker, Steven, Nowak, Martin A., and Lee, James L. 2008. ‘The logic of indirect speech’, *Proceedings of the National Academy of Sciences* 105(3): 833–838.
- Power, Camilla 1998. ‘Old wives’ tales: the gossip hypothesis and the reliability of cheap signals’, in Hurford, Studdert-Kennedy, and Knight (eds.), pp. 111–129.
- Reader, Simon M. and Laland, Kevin N. 2002. ‘Social intelligence, innovation, and enhanced brain size in primates’, *Proceedings of the National Academy of Science* 99(7): 4436–4441.
- Reboul, Anne 2014. ‘Live metaphors’, in Reboul (ed.), pp. 503–515.
- Reboul, Anne (ed.) 2014. *Mind, values and metaphysics: Essays in honor of Kevin Muligan*, vol. 2. Cham/Heidelberg/New York: Springer.
- Reboul, Anne 2015. ‘Why language really is not a communication system: a cognitive view of language evolution’, *Frontiers in Psychology* 24 Sept. doi: 10/3389/fpsyg.2015.01434
- Reboul, Anne and Moeschler, Jacques 1998. *La pragmatique aujourd’hui*. Paris: Le Seuil.
- Recanati, François 2004. *Literal meaning*. Cambridge: Cambridge University Press.
- Roberts, Craige 2004. ‘Context in dynamic interpretation’, in Horn and Ward (eds.), pp. 197–220.
- Scott-Phillips, Tom 2015. *Speaking our minds: Why human communication is different and how language evolved to make it special*. London: Palgrave MacMillan.
- Seyfarth, Robert M. and Cheney, Dorothy 2002. ‘What are big brains for?’ *Proceedings of the National Academy of Science* 99(7): 4141–4142.
- Skyrms, Brian 2004. *The stag hunt and the evolution of social structure*. New York: Cambridge University Press.
- Sperber, Dan, Clément, Fabrice, Heintz, Christophe, Mascaro, Olivier, Mercier, Hugo, Origgi, Gloria, and Wilson, Deirdre 2010. ‘Epistemic vigilance’, *Mind & Language* 25(4): 359–393.
- Sperber, Dan and Wilson, Deirdre 1995. *Relevance: Communication and cognition*. 2nd ed. Oxford: Basil Blackwell.
- Stanley, Jason 2007. *Language in context: Selected essays*. Oxford: Clarendon.
- Strawson, Peter F. 1950. ‘On referring’, *Mind* 59(235): 320–344.
- Számádó, Szabolcs and Szathmáry, Eörs 2006. ‘Selective scenarios for the emergence of natural language’, *Trends in Ecology and Evolution* 21(10): 555–561.
- Tomasello, Michael 2000. *The cultural origins of human cognition*. Cambridge, MA/London: Harvard University Press.
- Tomasello, Michael 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: The MIT Press.

- Tomasello, Michael 2009. *Why we cooperate*. Cambridge, MA: The MIT Press.
- Tomasello, Michael 2010. *Origins of human communication*. Cambridge, MA: The MIT Press.
- Tomasello, Michael 2014. *A natural history of human thinking*. Cambridge, MA: The MIT Press.
- Trivers, Robert 2002. *Natural selection and social theory: Selected papers of Robert Trivers*. Oxford/New York: Oxford University Press.
- Trivers, Robert 2011. *Deceit and self-deception: Fooling yourself the best to fool others*. London: Penguin.
- Villacencio, Aline and Poibeau, Thierry (eds.) 2017. *Language, cognition and computational models*. Cambridge: Cambridge University Press.
- Von Fintel, Kai and Matthewson, Lisa 2008. ‘Universals in semantics’, *The Linguistic Review* 25: 139–201.
- Whiten, Andrew and Byrne, Richard W. (eds.) 1997. *Machiavellian intelligence II: Extensions and evaluations*. Cambridge/New York: Cambridge University Press.
- Wilson, Deirdre and Sperber, Dan 2012. *Meaning and relevance*. Cambridge: Cambridge University Press.
- Wittig, Roman M. and Boesch, Christophe 2005. ‘How to repair relationships – Reconciliation in wild chimpanzees (*Pan Troglodytes*)’, *Ethology* 111: 736–763.

10 Genetic Biases in Language: Computer Models and Experimental Approaches

Rick Janssen and Dan Dediu

Abstract

Computer models of cultural evolution have shown language properties emerging on interacting agents with a brain that lacks dedicated, nativist language modules. Notably, models using Bayesian agents provide a precise specification of (extra-)linguistic factors (e.g., genetic) that shape language through iterated learning (biases on language), and demonstrate that weak biases get expressed more strongly over time (bias amplification). Other models attempt to lessen assumption on agents' innate predispositions even more, and emphasize self-organization within agents, highlighting glossogenesis (the development of language from a nonlinguistic state). Ultimately however, one also has to recognize that biology and culture are strongly interacting, forming a coevolving system. As such, computer models show that agents might (biologically) evolve to a state predisposed to language adaptability, where (culturally) stable language features might get assimilated into the genome via Baldwinian niche construction. In summary, while many questions about language evolution remain unanswered, it is clear that it is not to be completely understood from a purely biological, cognitivist perspective. Language should be regarded as (partially) emerging on the social interactions between large populations of speakers. In this context, agent models provide a sound approach to investigate the complex dynamics of genetic biasing on language and speech.

10.1 Introduction

10.1.1 Biasing Language

In this chapter, we argue not only that the best approach to understanding the origins and present-day diversity of language is rooted in evolutionary

theory, but also that extra-linguistic factors, more specifically biological ones in our genes, may play an important role in shaping language. Likewise, these factors do not act in a void, but interact with multiple constraints and affordances on different scales in parallel. So-called *cultural evolution of language* (Section 10.1.2) must thus be seen in a rich context (partially) molded by the biological and cognitive entities that ultimately acquire, use, and transmit language – us. Important factors in this context are therefore represented not only by the brain – it has been recognized for a while now that the brain indeed shapes language (Christiansen & Chater 2008) – but also by the anatomy and physiology of the vocal tract and hearing organs. Just to illustrate, it has been recently suggested (Butcher 2006) that the very high rates of chronic otitis media (an infection of the middle ear that impacts hearing) affecting Australian Aboriginal children might explain striking features of the phonological systems of the Australian languages such as a lack of fricatives and many distinctive places of articulation. This process of *biasing*, whereby extra-linguistic factors, ultimately with a genetic basis, can affect cultural evolution of language, has been suggested to be a rather general influence playing a role in explaining not only universal tendencies (when these biases are shared across the whole human species) but also linguistic diversity (when the biases differ between human populations in magnitude or direction) (Ladd, Dediu & Kinsella 2008; Dediu 2011).

There are multiple lines of evidence supporting genetic biasing of language and many interesting directions to explore,¹ but we focus on a very specific set of questions: What can we conclude about the nature and effects of such biases from the body of computational modeling work and experimental approaches (using both human subjects as well as animal models) on language change and evolution? To this end, we begin by discussing some fundamental notions necessary for a cultural evolutionary approach and the influence of genetic biases (Section 10.1.2), followed by an overview of some relevant computer models (such as various iterated learning approaches [Sections 10.2.1 and 10.2.2] but also models that do not belong to this tradition [Section 10.2.4]). Computational models such as these are particularly interesting to investigate processes that develop on long time scales, since they allow for experimental manipulation not available otherwise (de Boer & Fitch 2010). However, a sound empirical foundation, possible in series with other models, is hereby essential in order to make testable predictions, such as those corroborated by the experimental results in Section 10.2.3. Finally, we discuss models that address the possibility

¹ Both authors are part (together with Dr. Scott Moisik) of the **Genetic Biases in Languages and Speech** (G3bils) project, which actively investigates the influence of the vocal tract as a biasing factor on phonetics and phonology (<http://www.mpi.nl/departments/language-and-genetics/projects/genetic-biasing-in-language-and-speech>).

of feedback from culture into the genome through the Baldwin Effect (Section 10.3). This overview of a diverse literature (and we must highlight the fact that few studies were designed with genetics in mind) suggests that while genetic biases can indeed affect language, it is still too early to draw any general conclusions regarding the strength required for such biases to become manifest and be measurable in human populations (Section 10.4).

10.1.2 Cultural Evolution of Language

Many accounts of the nature, origins, and evolution of language do not consider evolutionary processes to play any important role (e.g., Chomsky 1986). Even the field of historical linguistics (e.g., Campbell & Poser 2008), which tries to understand the factors, processes, and outcomes of language change across time, does not have an evolutionary outlook and seems generally rather critical of approaches that use such concepts and methods (see, for example, the quite cold reception of modern phylogenetic approaches to questions of language relatedness, the dating of proto-languages, and the expansion of language families such as Indo-European and Austronesian (e.g., Gray & Atkinson 2003; Pagel, Atkinson & Meade 2007; Dunn, Greenhill, Levinson & Gray 2011; Bouckaert, Lemey, Dunn, Greenhill, Alekseyenko, Drummond, Gray, Suchard & Atkinson 2012)). On the other hand, there are other proposals that consider biological evolution to be the main explanatory factor behind the human use of language (e.g., Pinker & Bloom 1990), but they miss the intervening causal role played by cultural evolution by emphasizing biological nativism.

The emphasis on biological evolution is not surprising. Darwinian theory, based on the principles of replication, variation, and selection, has proven to be an immensely powerful approach to explain biological complexity (Carroll 2005). In a nutshell (and glossing over many fascinating aspects of evolutionary biology), when a population of organisms reproduces, slight variations will be introduced in the offsprings' genome by mutations. These mutations are essentially random, most of them having a neutral or negative effect on an organism's phenotype. Mutations might, for instance, cause inheritable diseases such as sickle-cell anemia (OMIM² 603903) or developmental speech dyspraxia (OMIM 602081). A small number of mutations, however, might have positive effects. They might, for example, give an animal a slightly increased resistance to certain pathogens, enlarged cardiovascular capacity, or enhanced cognitive capabilities. If these improvements are small, just one of them is of course unlikely to be saliently noticeable. However, advantageous mutations

² For the sake of brevity, we will refer to OMIM (Online Mendelian Inheritance in Man; <http://omim.org>) unique identifiers that give access to brief up-to-date descriptions and the relevant literature.

will accumulate as an effect of selection whereby the organisms with an increased fitness (in part ascribable to these advantageous mutations) are more likely to reproduce, transmitting the mutations to their offspring.

A common misconception is that evolution is teleological (Hanke 2004). For instance, the notion that we, *Homo sapiens*, have evolved to have large cognitive capabilities is most likely a matter of circumstances more than anything else.³ As a matter of fact, natural selection should be viewed as a system merely acting as a filter on the existing variation in a population, in many ways similar to many optimization algorithms used in computer science (e.g., metaheuristics). Thus, evolution produces ad hoc solutions that are merely appropriate for the situation at hand, without any notion of design, aesthetics, elegance, rational insight, or intentionality. This lack of foresight is strikingly apparent when looking at what could be considered “design errors” such as photoreceptors pointing away from the direction light strikes the retina in vertebrates.⁴ Design errors demonstrate that showing that Darwinian processes are at work does not warrant the conclusion that they should necessarily result in sophistication and, conversely, that if we observe sophistication, Darwinian processes are per definition responsible.⁵ Nevertheless, natural selection is widely considered the most powerful explanation for the origin of biological complexity on phylogenetic time scales (i.e., pertaining to the formation of taxonomic groups). As such, the proposition that biological evolution is responsible for the complex system of language as well seems to be a logical one (e.g., Pinker & Bloom 1990). Upon closer consideration, however, given the fact that languages change much faster than any biological evolution to fully account for it (Christiansen & Chater 2008), this appears unlikely.

Even though the principles of replication, variation, and selection are simple, biologists have been vigorously debating the “level” they act on. Intuitively, this level might appear to be located at the scale of individuals, i.e., organisms competing with each other for food and mates. The idea of group-level selection on the other hand has been invoked to explain some widely observed behaviors such as altruism as demonstrated by, for instance, the use of distress calls in

³ There is however considerable debate on the evolution of the human brain, ranging from an effect of social (Dávid-Barrett & Dunbar 2013), sexual (Miller 2001), environmental (Calvin 2002), or other selective pressures.

⁴ There is some debate on whether the inverted vertebrate retina is the result of a historically frozen maladaptation or a trade-off between optical and other physiological (e.g., metabolic) costs (Kröger & Biehlmaier 2009).

⁵ What is described here is also known as the logical fallacy of “affirming the consequent”. Although natural selection is one mechanism that explains complexity, other mechanisms might do so comparably well. For instance, ice-crystal or spiral-galaxy formation does not require descent with modification but self-organizes following mere physical, *in situ* interactions (e.g., Lin & Shu 1964).

groups of meerkats when spotting a predator (Wynne-Edwards 1962; Wynne-Edwards 1986).⁶ More recently, the gene-centered view of selection has been popularized, providing an alternative for explaining altruism and introducing the concept of the *extended phenotype* that reaches beyond the confines of the biological organism and includes “expressions” such as bird nests and termite mounds (Hamilton 1963; Williams 1966; Dawkins 1976).

Essentially, the debate on the level of selection centers on the conceptualization of the replicating unit (or *replicator*) that drives evolution. Interestingly, if Darwinism is not confined to the domain of biology alone (and why should it?), these replicators might not just exist at different levels of a biological system (multilevel selection; Okasha 2006; Wilson & Wilson 2008) but also in different domains altogether. One such domain might of course be human language. So, what might a linguistic replicator look like?⁷ Consider that a human community engaged in linguistic exchange is producing a population of utterances that are actively used in everyday speech. Similarly to how organisms are “composed” of genes, utterance can be regarded to be composed of *linguemes*, “building blocks” that can change and be recombined to form new utterances (Croft 2000). Examples of these linguemes – linguistic replicators – are phonemes, morphemes, words, or even grammatical rules. While genes are transmitted during reproduction from parent to offspring, linguemes are transmitted from teacher to learner, for instance during, but not restricted to, first language acquisition during childhood. As we know, genes that are likely to be transmitted during reproduction will, *ceteris paribus*, eventually spread throughout a population. Similarly, those linguemes that are likely to be transmitted from teacher to learner will eventually proliferate in a cultural population of linguistic units. Even though there are obvious differences between genes and linguemes, such as the degree of horizontal information flow, the encoding medium, and the speed of change and transmission noise,⁸ there are the (striking) common principles of reproduction, variation, and selection (Levinson & Gray 2012) that seem to provide a valid explanation for the complex phenomenon of language without having to rely on biological determinism alone.

⁶ The confusion on the level of selection is also visible in the abuse of Darwinian theory in the justification of e.g., eugenics and racialism during the early 20th century.

⁷ Even if this question remained unanswered, it would not invalidate the viewpoint that language evolves in a cultural domain. Mendelian genetics has been successfully practiced before the discovery of the encoding medium, DNA. Our current understanding of cultural evolution might be of a comparable advancement.

⁸ The term “noise” should be understood in its broadest sense here. The brain has an active, heuristic role in selecting which linguemes are transmitted (Christiansen & Chater 2008). Noise in cultural selection is therefore not as uniform as in genetics but likely structured to some degree (see e.g., Farrell, Wagenmakers & Ratcliff 2006 for the presence of pink noise – structured noise showing self-similarity – in cognitive performance).

Still, we have not discussed a good analogue of selective pressure in biology. Section 10.2 addresses this issue by explaining selection on linguemes as a *transmission bottleneck* by means of a series of computer modeling experiments. Interestingly, many of these models imply, as briefly mentioned in Section 10.1.1, that cultural evolution of language converges to particular states and that this convergence is influenced by biasing factors. Examples of these factors could be environmental (e.g., altitude, humidity) (Everett, Blasi & Roberts 2015), social (e.g., peer pressure, conformity bias, sexual selection), cognitive (e.g., working memory capacity, language impairments like aphasia), genetic (e.g., *FOXP2* (OMIM 605317), *ASPM* (OMIM 605481), *MCPH1* (OMIM 607117)), or anatomical (e.g., hard-palate curvature, jaw length) (Henrich & McElreath 2003). Biases such as these might co-exert a likely very subtle selective pressure in cultural evolution of language, potentially providing an explanation for why languages differ. Importantly, these differences do not imply strong biases per se: It is very unlikely that human languages differ because populations have different innate, all-or-nothing cognitive or anatomical predispositions that allow for or prohibit certain language features. Instead, genetics might, for example, influence a speaker's "effort" associated with producing certain speech sounds or patterns. These efforts might then, over time, lead to a cascading effect via cultural evolution. In such a case, even the weakest bias might get amplified to the point where it could (even) lead to phonemic discriminability (Ladd et al. 2008; Dedić 2011). In fact, this amplification effect has been observed in computer simulations (Section 10.2.2) and, arguably, in animal models (Section 10.2.3).

As a slightly more concrete example of anatomical biasing factors, it is not hard to imagine a particular hard palate shape that, say, eases ingestion and disturbs the production of particular speech gestures while facilitating others. While speculative, similar propositions have been postulated before. For example, Ladefoged (1984) made the observation that while Italian and Yoruba (a Niger-Congo language) have very similar vowel systems, there are nevertheless slight differences, possibly due to a minor anatomical between-group variance (e.g., a slightly larger mouth opening in Yoruba speakers, lowering the second formant in vowels). Importantly, such a very subtle, almost undetectable bias, would not prevent any Italian from speaking Yoruba if being exposed to it from infancy (and the other way around). On a *glossogenetic* level (i.e., at the time scales concerning historical language change), however, these subtle anatomical biases might become saliently expressed on a cultural, population level. Culturally amplified biases such as these could explain the present-day distributions of features seen in human language, such as Yoruba second formant lowering or word order, not as having arrived abruptly by a chance mutation giving rise to "language modules" (Fodor 1983) or "language acquisition devices" (Chomsky 1965), but as (weakly) emerging from the interactions between large

populations of situated speakers. This then effectively describes a *dynamical system* that slowly gravitates toward particular attractors in an articulation-landscape that is (partially) formed by these biases.

To summarize, language can thus be seen as a system evolving in itself, thereby reducing the plausibility of strong biological, even nativist explanations for its structure, diversity, and evolution. However, this does not imply that extra-linguistic factors are of no effect. When we consider the two evolving systems – biological and cultural – in parallel, it is not hard to see how they might be compared to how organisms, such as predator and prey or parasite and host, are *coevolving* (Richerson & Boyd 2008). Likewise, culture and biology might be exerting reciprocal selective pressure on each other, raising the possibility that, on one hand, biology influences (or biases) culture and, on the other, stable cultural features become “inscribed” into the (biological) genome (Section 10.3). Even without addressing the full complexities of coevolution (Section 10.3), we can already be sure that cultural and biological evolution can by no means be considered independent from each other.

10.2 Computer Models of Cultural Evolution

10.2.1 Iterated Learning

As discussed in Section 10.1.2, cultural evolution of language centers on the idea that linguistic replicators are in competition with each other, similarly to how genes might be in biology. To investigate cultural evolution in computer simulation, Kirby & Hurford (2002) developed the Iterated Learning (IL) agent model. IL is of particular interest to the topic of genetic biasing not only because it explains some features unique to human language (Hockett 1960) without any (*a priori*) requirement on biological evolution but also because it allows for a precise specification of the nature and strengths of these biases (Section 10.2.2).

Briefly summarized, IL simulates language transmission from teacher to learner. This works by agents conveying “meanings” to each other by transmitting “utterances.” At the time of this writing, there are many incarnations of the procedure. To start of simply, in Kirby & Hurford (2002), only one pair of teacher and learner exists at any time (a population of two), where the old learner replaces the teacher when a new learners is introduced. This population model is now often considered to be a sort of sequential teacher-learner chain (Fig. 10.1). The conversion of meanings to utterances (and vice versa) is determined by the agents’ internal rules (i.e., a “grammar”).

In the chain model, once all conveyed meanings are learned, the learner will take the role of teacher and attempt to convey the interpreted meaning to the next learner. By transmitting meanings from agent to agent, language transmission, and thereby cultural evolution of language, is modeled. The longer

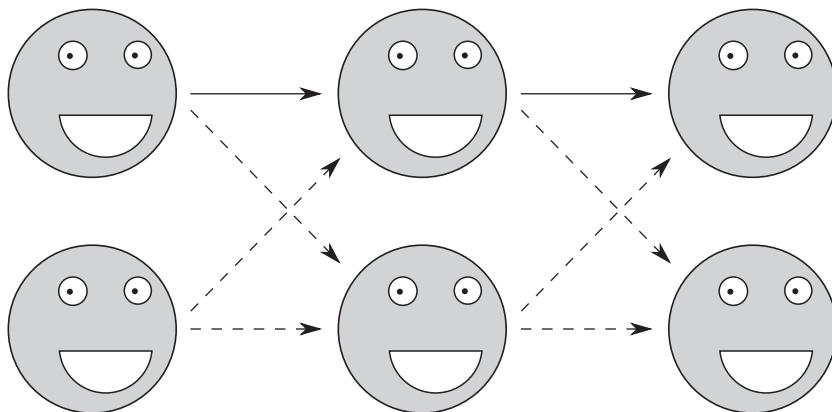


Figure 10.1 The IL social structure. Agents transmit utterances to each other following the arrows. Solid arrows mark those channels available in the monadic experiments, modeling vertical information transfer (e.g., from single parent to single learner). Dashed arrows mark communication channels available in polyadic setups, modeling oblique flow (e.g., teachers that teach multiple learners, while learners learn from multiple teachers). The classical IL paradigm does not normally allow for horizontal communication (e.g., between agents from the same “generation”). Note that there is no biological evolution in IL: agents’ architecture remains unchanged.

the chain, the more transmissions occur and the larger the time-frame that is modeled. Importantly (as we will see at the end of this section), the number of meanings that have to be, in principle, expressible is much larger than the number of permitted utterances transmitted between teacher and learner.

When having to produce an utterance, meanings to be conveyed are selected randomly from a fixed, global pool and structured following an “Agent-Patient-Predicate” syntax. For example, a meaning could be structured as $\langle \text{Agent} = \text{Henk} \rangle$. When a teacher has no specified utterance associated with a meaning to be expressed, a new utterance has to be invented. This is done by generating arbitrary-length, random character substrings for the components in the meaning that are unknown, while the known components are simply filled in. For example, if a teacher would utter $\langle \text{Patient} = \text{Ingrid} \rangle$ as “ingrid,” $\langle \text{Predicate} = \text{Kust} \rangle$ as “kust,”⁹ but have no way to express the meaning $\langle \text{Agent} = \text{Henk} \rangle$, the final utterance might become something as “fmguhba kust ingrid.” Agents are initially naive. Therefore, when starting the simulation, all utterances produced by the first agent in the chain are completely random.

⁹ Dutch third-person singular for “to kiss.”

Learners do not simply internalize what they hear when they get exposed to utterances, but they are able to make generalizations. The precise details of these generalizations can vary from study to study, but what is more important is that they all lead to similar long-term dynamics as meanings get conveyed from agent to agent over long time periods. To summarize, all strategies follow the principle of substituting many complex rules with fewer simpler ones, when there is similarity between the complex rules. More precisely: when multiple meanings are expressed by similar utterances (i.e., they have similar syntax subtrees), a new rule is generated that substitutes the lower-level syntax instance with a more general, higher-level rule. This rule is then applicable to many meaning-to-utterance conversions. These more general, higher-level rules thus enable the agent to generate *compositional* syntax.

Having laid out this framework of cultural evolution of language, the results provided some interesting observations. Over many transmissions, the size of the grammar (i.e., the number of rules converting meaning to utterance) and the number of expressible meanings were recorded. These measurements showed three distinct phases (Fig. 10.2).

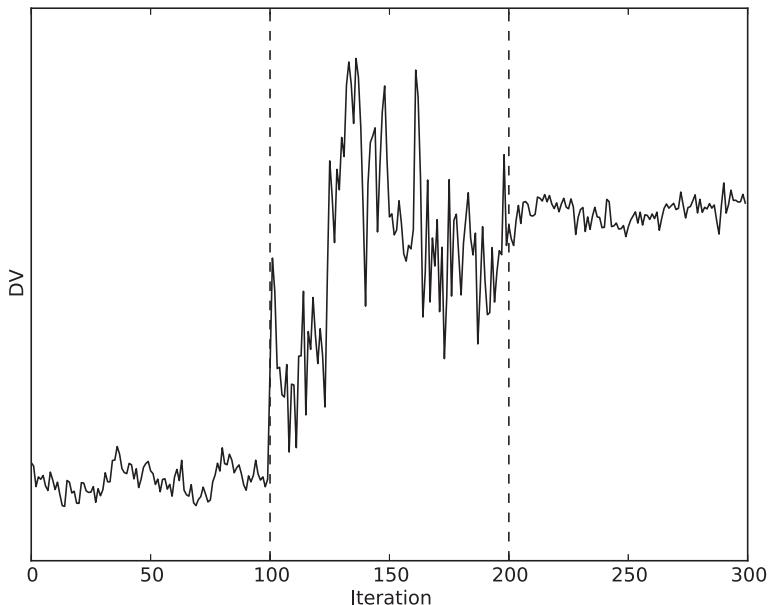


Figure 10.2 A generalization of the phases observable in a typical IL run. Shown is some dependent variable (DV), such as the number of expressible meanings, over time. The second phase (Iteration 100–200) shows a chaotic phase transition following a first phase of stabler dynamics, in turn leading to a state of (semi-)stationary convergence.

In the first phase, both the number of expressible meanings and the size of the grammar remain small, with minor fluctuations over time. Inspection of the agents' syntax trees show that they are completely flat. This is also shown when inspecting the actual utterances used by the agents. These are completely arbitrary and random. For example, the meaning $\langle \text{Agent} = \text{Henk}, \text{Patient} = \text{Ingrid}, \text{Predicate} = \text{Kust} \rangle$ could be expressed by the utterance "ddadbbbabeedae," while a similar sentence $\langle \text{Agent} = \text{Sjaak}, \text{Patient} = \text{Ingrid}, \text{Predicate} = \text{Kust} \rangle$ could be expressed by the utterance "d." In other words, utterances are completely idiosyncratic. Each meaning is coupled with a unique utterance and vice versa, and there is no underlying, general structure to them. In many ways, this first phase can be likened to a *proto-language* as hypothesized to be used by earlier hominids using idiosyncratic grunts or gestures to convey meaning (Dediu & Levinson 2013).

The second phase marks a period of large fluctuations in grammar and meaning size, following a brief burst of rapid inflation. It appears there now exist syntactic categories. Some meaning components might now be regularly expressed by the same set of characters, although this is only true for a small number of them. This is reflected in the syntax trees the agents use, which are no longer completely flat, but show occasional branching (the language is said to be partially compositional). Overall, due to the large fluctuations, it is hard to understand what is precisely going on in this phase. One could argue, from an optimization algorithm perspective, that the agents are exploring a search space, briefly occupying some local optima, none of them stable enough to lead to terminal convergence. From a dynamical systems perspective, this second period is typical of what is known as a *phase transition*, analogous to phase transitions in physical systems, e.g., freezing and melting of liquids and solids, respectively.

Eventually however, the chaotic fluctuations in the second phase settle down in a third, static phase where all meanings are expressible using a small number of grammatical rules. In this phase, the language is largely compositional. The rules the agents use are applicable to many linguistic instances and syntax trees reflect this by showing extensive branching of syntactic categories. Few meanings are any longer expressed idiosyncratically. For example, the meaning component $\langle \text{Agent} = \text{Henk} \rangle$ could be consistently expressed by the substring "qasd." Furthermore, the usage of particular types of meaning components (i.e., verbs or nouns, predicates or agents/patients) is reflected in the regular positioning in utterances. This is a clear analogue to the use of a consistent word order in actual human languages. Kirby & Hurford (2002) further illustrated the potential of the IL model by showing emergence of *recursive* syntax, but now following a simplified model with agents embedded in a linear chain. Here, utterances are propagated through the chain, with agents first taking the role of learner followed by subsequently assuming the role of teacher

(Fig. 10.1). In this experiment, utterances were interpreted by a perceptron (a simple type of feedforward artificial neural network), while they were produced using statistical inferencing. Interestingly, in an additional experiment (Kirby & Hurford 2002), the ratio of requested meaning was changed such that some meanings were either often or rarely requested, mirroring a Zipfian distribution (Zipf 1949). In this case, the commonly requested meanings remained idiosyncratic, resisting the transmission bottleneck induced generalizations. This parallels the use of irregulars, like the verb “to be” in English.

How might the emergence of compositionality, and even recursive syntax, follow from simple iterated transmission? Consider how agents need to transmit a large number of meanings, with only a limited number of opportunities. This information bottleneck implies that there is no “time” for learners to exhaustively transmit every meaning if they were expressed by idiosyncratic utterances. In such a case, old idiosyncratic utterances are “forgotten,” continuously requiring the invention of new ones. However, if, instead of idiosyncratic utterances, compositional ones could be transmitted, many more meanings would “fit through” the bottleneck. This is because such utterances are based on a more general grammar, applicable to many meaning instances, instead of on an idiosyncratic grammar, which only describes very few meaning instances. Only the most general rules therefore survive the selective pressure the information bottleneck exerts, because only those general rules, applicable to many linguistic instances, are invoked often enough to be guaranteed transmission. Using Darwinian terminology: the more general linguistic replicators have a larger chance of being summoned, thus they are “fitter,” and thus they are more likely to actually replicate when compared to idiosyncratic replicators. This explains how the complex system of human language can be considered as itself evolving to become learnable, instead of the agents, e.g., human beings, evolving to learn the language. This way, it appears cultural evolution alone can be sufficient for features unique to human language, such as compositionality (Hockett 1960), to emerge.

Altogether, however, one might question the claims of the IL experiments that compositionality and recursion in the resulting languages are the effect of the transmission bottleneck alone. First of all, instead of becoming compositional, a language might equally well converge to a state where every meaning is expressed by a single, simple utterance. In this case the language has become maximally transmissible (learnable) but has not become expressive – the language is essentially one big homonym (this issue will be further discussed in Sections 10.2.2 and 10.2.3). Second, the learning algorithm agents use, whether based on inductive logic, neural networks, Bayesian inferencing (Section 10.2.2), or on some other method, is an essential component of the model that co-determines what attractors the language will converge to. For example, in the case of neural networks, it has been shown that only particular

types of networks lead to this convergence, while others cannot overcome an initially naive state and initiate autocatalysis, or are unable to produce faithful reproductions in noisy environments (Smith 2001). This latter criticism then makes a good case that the IL experiments do not explain glossogenesis – the (evolutionary) origin of language in our species from a non-linguistic state. In effect, IL “merely” explains how e.g., the cognitive apparatus, once it provides adequate functionality, is able to shape the convergence of cultural evolution, while it says little on how this apparatus was evolved, whether it was specifically tuned for language learning by domain-specific adaptation, or whether human agents apply more general learning strategies to communicate with each other (explaining the linguistic features we see today as a result of domain-general *exaptations* – the utilization of existing adaptations already in place for novel purposes). These, in many ways even more theoretical, questions will be further addressed in Section 10.3. For now, it can be said that the IL framework mainly models language change on a glossogenetic level, while *ontogenetics* (i.e., developmental factors that shape the learning algorithm) are of a significant influence but not the main focus.

Now that we have established that the learning algorithm, modeling human cognition, has a large influence on the convergence of language via cultural evolution, we might propose a generalization. Instead of the human brain exerting biases, we could imagine, as mentioned in Section 10.1.1 and 10.1.2, that human anatomy and physiology, even at a relatively low, mechanistic level, might exert such biases comparably well. For example, the flexibility of the tympanic membrane or the tone of muscles attached to the ossicles in the inner ear might impose low-level perceptual biases. Alternatively, during speech production, it is not hard to imagine how the shape of the vocal tract, such as hard-palate curvature or lower vocal tract volume, could make the production of particular speech sounds easier or harder, exerting biases in their own terms. As theorized (Section 10.1.2), biases such as these might get saliently expressed even when very small because they can be amplified when iteratively transmitted from speaker to learner. However, bias amplification rest on models with strong assumptions, the *Bayesian IL* models, addressed in Section 10.2.2.

10.2.2 Bayesian Iterated Learning

The IL models discussed in Section 10.2.1 explain human language as a system emerging from the interactions between agents, strongly implying that commonly seen linguistic features need not be biologically innate or had to evolve by natural selection. However, a series of follow-ups on the IL experiments casts doubts on these assumptions, while making interesting predictions in their own right. The IL models were based on agents using various algorithms such as ones based on neural networks or inductive logic. However, Griffiths and

Kalish (2007) use agents that use Bayes' theorem (Eq. 10.1) to reason about language features.

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h \in H} P(h)P(d|h)} \quad (10.1)$$

In the context of IL, Bayes' theorem (Eq. 10.1) describes how agents derive a distribution of language hypotheses (called the *posterior* distribution, or $P(h|d)$) from the observed data (in the form of a *likelihood distribution*, $P(d|h)$) and some sort of bias (a *prior distribution* on language hypotheses, $P(h)$). This bias could, for instance, represent the neural, anatomical, or genetic biases discussed in Section 10.1.2. (The denominator in Eq. 10.1 denotes a normalizing factor.)

To provide a concrete illustration of how this works, consider an IL chain as described in Section 10.2.1, but now with Bayesian agents transmitting utterances to each other (Fig. 10.1). Suppose an agent would have to decide on word order in the utterances they produce, Subject-Object-Verb (SOV) or Subject-Verb-Object (SVO). If the agent's prior distribution would be uniform, i.e., express a 50% preference for either SVO or SOV, the posterior distribution would be completely determined by what word order is used by a teacher agent, and how likely this data is expected to be under different language hypotheses (for example, perceiving an SVO sentence could imply a 95% likelihood that the teacher agent generated it under an SVO hypothesis). Conversely, if this likelihood distribution would be uniform, the prior would dominate. Of course, these simplified situations would not normally occur and the posterior distribution on language hypotheses would always be a product of both likelihood and prior.

Once a posterior distribution has been established, there still remains the issue of what hypothesis an agent will select. To simplify, two options are available: sampling and maximizing. When agents are sampling, language hypotheses are selected with a chance proportional to their probability in the posterior distribution. When maximizing, the hypotheses with the largest posterior is selected. Interestingly, the selection strategy used has a large effect on the resulting languages. When sampling, languages converge on a (Markov chain's) stationary distribution that exactly mirrors the prior. This has the implication that, if human agents are equivalent to Bayesian samplers, the languages we see today strongly reflect biologically innate predispositions. When maximizing, however, convergence is less well understood, but largely seems to reflect only the *ordering* of the hypotheses in the prior distribution. For example, if the prior distribution would express an 80% bias for SVO word order and a 20% bias for SOV, we would expect SVO to be used more often than SOV, but likely not following the 80–20 ratio.

Kirby, Dowman and Griffiths (2007) emphasized the amplification of weak priors by maximizer populations, i.e., a prior with only a slight edge over the other priors might come to dominate over time. This would imply that strong effects (e.g., common speech patterns of phonemes, “universals”) would not require strong biases (e.g., neural modules). Smith and Kirby (2008) then demonstrated that a population of maximizer agents is the evolutionary stable one over a sampler population, i.e., a maximizer population will resist invasion by a sampler minority, but not the other way around. The reason for this is that maximizers are more certain what other agents are doing (i.e., maximizing, selecting the most likely language hypothesis, just like themselves) while with samplers there is a larger chance agents will choose different hypotheses. Since we know that maximizing is the evolutionary stable strategy, and since we know that maximizing can amplify biases, the authors conclude that, altogether, selection should be neutral with regards to prior strength. In other words, cultural evolution leads to *shielding* of bias strength (Kirby et al. 2007).

While intriguing, we must emphasize that these conclusions make a number of assumptions that could be questioned. First, there is the assumption that a bias’s strength is conspicuously apparent to listeners, and that they all are comparably salient (in other words, that they all have an effect of comparable magnitude on perceived utterances). But we know that many high-level language features are invariant to lower-level details (e.g., allophones). For instance, a large bias that induces prohibitive effort on producing the voiced bilabial stop [b], will have no effect on phonemic intelligibility in Spanish if the biased speaker would simply shift to a voiced labiodental fricative [v]. Second, it is assumed biases have homogeneous peripheral costs, i.e., they are all similar in their effects on an organism’s fitness besides the communicative one. But if a particular weak bias is disproportionately costly with respect to, for instance, the ability to swallow or breathe, that bias is probably still selected against. Thus, the notion of selectively neutrality only applies to the bias’s effects on communicative accuracy. Third, IL assumes that not only the utterances but also associated meanings are observable by agents. However, it is less clear how “meanings,” particularly more abstract ones, can be observed by human speakers. Fourth, it is known that the apparent dichotomy between maximizers and samplers Bayesian IL postulates in reality follows a continuum (Kirby et al. 2007); agents can occupy a position that is intermediate between sampling and maximizing. Such agents (that are thus not perfect samplers) over time approximate the behavior of maximizers, causing the eventual language distribution to mirror only the prior ordering. Finally, it is assumed that language transmission can be likened to vertical transmission in a linear, monadic chain (i.e., each learner is being taught by exactly one teacher) or to a population of infinite size (Griffiths & Kalish 2007) (Fig. 10.1). When situated in heterogeneous, polyadic chains (e.g., learners having two or more

teachers), however, language convergence strongly diverges from the monadic chain behavior.

In Smith (2009), it was shown that a polyadic chain of sampling agents converges to the language that has the largest prior, while languages with weaker priors get suppressed. This runs contrary to the monadic results that show samplers converging on a distribution that exactly mirrors the prior, suggesting a type of conformist dynamic. Furthermore, increasing bottleneck size increases transmission fidelity and promotes convergence on the strongest prior (contrary to a monadic chain, where a smaller bottleneck leads to faster prior expression (Kirby et al. 2007)). Similar deviating results have been shown by Ferdinand and Zuidema (2009). First, they showed that homogeneous polyadic maximizers behave almost the same as monadic maximizers do, while polyadic samplers no longer converge to the prior distribution. This is explained by the observation that a learner might receive data as a product of multiple teacher agents that might have entertained different language hypotheses. In other words, the data learners receive is generated from a “virtual” distribution they have no explicit internal representation for, therefore they can no longer be Bayesian rational. The study also investigated heterogeneous polyadic chains where learners had different prior distributions (e.g., teacher agents had different biases). When agents were sampling, the language eventually reflects the average of multiple, homogeneous sampler runs. When maximizing, things get more obfuscated, as language convergence codepends on prior strength and likelihood structure and distinctiveness.

In Dediu (2008, 2009), complete populations of spatially dispersed agents were investigated, thereby including horizontal information flow instead of the purely vertical or oblique transmission seen in the standard IL chains (Fig. 10.1). Dediu (2008) showed that (in populations of non-Bayesian agents) two kinds of biases had different effects on language convergence. With the “initial expectation” bias (e.g., an innate predisposition for some language features), these biases very soon get overruled by linguistic drift. However, the “rate of learning” bias (e.g., an adaptive tendency to acquire particular language features) behaves more akin to Bayesian monadic samplers in the sense that weak biases get amplified through cultural transmission. Second, in Dediu (2009), actual Bayesian agents were then used in the population model, which showed that agents (whether samplers or maximizers) behave as monadic chains of samplers, contradicting the results from the polyadic chains from (Ferdinand & Zuidema 2009; Smith 2009).

These anomalies resulting from relaxing assumptions imply that care must be taken when generalizing findings obtained from Bayesian IL models. The findings in Dediu (2008, 2009), Ferdinand and Zuidema (2009), and Smith (2009) give strong testament to the issue of how to interpret the notions of Bayesian maximizing and sampling agents and to what extend human agents behave as

them. Even in the simple, monadic Bayesian chains introduced by Griffiths and Kalish (2007), it is hard to generalize conclusions on samplers and maximizers to human “agents” since we cannot assume human beings are Bayesian rational in the first place (Ferdinand & Zuidema 2009). Given this insight, it seems that the claim that cultural evolution of language converges on a distribution that mirrors speakers’ innate biases is definitely too strong. This is further highlighted by the results that show that different social topologies (vertical or oblique monadic or polyadic chains, or horizontal populations of agents) can result in different outcomes of the convergence process, and these results sometimes even contradict each other. Such apparent paradoxes, however, can possibly be explained by noticing that the experiments use, for instance, slightly different learning algorithms. As we have seen, a similar sensitivity is seen in the population architecture the IL models follow, and the question which structure is most powerful and most realistic remains unanswered (Mesoudi & Whiten 2008). We note that this sensitivity to initial conditions is typical of complex systems, implying that one cannot explain cultural evolution of language using only a reductionistic, component-based account, but that one, at some point, needs to include the interactions between said “components.” This leads us to raise the question to what extend predictions made by reductionist models such as Bayesian IL are corroborated by experiments on human subjects, which is discussed in Section 10.2.3.

Many of the anomalies of the Bayesian models could potentially be addressed by extending the hypothesis space the agents use in some way or another. For example, Burkett and Griffiths (2010) modified their original model to allow for multiple, heterogeneous teachers by using a hyperprior (a prior distribution over nested prior distributions; Bernardo & Smith, 2009). Using the same approach, Smith, Tamariz, and Kirby (2013) explicitly showed that compositionality in language depends not only on a requirement for learnability or generalizability but also on expressivity (a concern we raised in Section 10.2.1 and that is also observed in IL studies on human subjects; see Section 10.2.3). Using these modifications, it is theoretically possible to model any cognitive process by using nested hypothesis spaces in a hierarchical configuration (Perfors 2012). However, we argue that this more complete specification of the hypothesis space, powerful as it might be, negates one of the most appealing aspects of Bayesian IL, namely its simplicity and its tractability. In these cases, the use of Bayesian agents, in terms of epistemological comprehensibility, reverts to paradigms that are often regarded as opaque, such as artificial neural networks.

To conclude, the Bayesian IL models give us some strong suggestions on how genetic biases might be expressed in languages, but one should keep in mind that they are strongly reductionist in nature, while describing a system that shows all the characteristics of being dynamical. The prediction that genes

might be shielded from natural selection following cultural evolution of language should therefore probably not be taken as the final word on this topic, although it is an intriguing one nonetheless (see Section 10.3 for another example of shielding). Probably the main take-home message from these studies is that genetic biases can indeed affect the outcome of cultural evolution of language, but in very complex ways, and that bias strength is probably not linearly related to its outcome.

10.2.3 Human Subject Trials and Animal Models

One obvious way of validating the IL models is by testing if human subjects perform in a similar manner. A follow-up on the computational IL studies employed chains of human “agents” that had to generate and transmit utterances (strings of characters) describing meanings (pictograms that represented colored, moving objects, e.g., a blue, spiralling square) (Kirby, Cornish & Smith 2008). Utterance-meaning pairs were then relayed to a second subject tasked with replicating them, in the process transmitting them to a third learner, etc.

Results showed that the “language” the subjects produced converged to an underspecified state (i.e., many homonyms were used), paralleling some of the computational IL experiments mentioned in Sections 10.2.1 and 10.2.2 by Smith et al. (2013). However, when homonyms were filtered in the transmission line, the utterances became much more expressive while remaining learnable. Again, mirroring the computational IL models, utterances became increasingly structured (i.e., compositional) further down the IL chain. Moreover, since, from the subjects’ perspective, both experiments were of an indistinguishable nature, it was concluded that subjects’ intentions, learning strategies, or linguistic background (e.g., their native language) were not a factor in the outcome. However, we speculate that this latter conclusion (on linguistic background) is based on using homogeneous subject pools with little variation. More specifically, if pools were used that grouped subjects on cultural or ethnic background, language convergence is more likely to be different between groups. If this hypothesis indeed was to be established, it could suggest possible innate/acquired biases.

Similar studies on human subjects (Perfors & Navarro 2011; Perfors & Navarro 2014) showed comparable results, but instead of using pictograms, stimuli consisted of simple squares of different sizes and colours. Multiple trials were conducted where the stimulus space followed a smooth gradient (i.e., stimuli each differed from each other to a similar extend) or a more discontinuous one (i.e., some stimuli were very similar, while others were very different). In the gradient condition, utterances converged to an underspecified state, similar to the pictogram stimuli without the expressivity requirement by Kirby et al.

(2008) and what was emphasized in the Bayesian model by Smith et al. (2013). With the discontinuous stimuli, however, the language used at the end of the IL chain reflected the discrete profile of the stimulus-space, even without any explicit requirement for expressivity. In other words, language convergence is influenced not only by innate (e.g., genetic, neural, anatomical) biases but also by the environment (represented by the stimuli distribution) subjects were situated in. A discontinuous environment (i.e., one that provides, one could say, a template for semantic categorizations and an implicit requirement for expressivity) is then able to prevent language convergence to an underspecified state. Furthermore, while the Bayesian sampler studies predict faster convergence with smaller bottlenecks (Section 10.2.2), Perfors and Navarro (2014) show that a larger bottleneck will reflect the environmental biases more strongly. Indeed, a small bottleneck leads to another instance of convergence to an underspecified state (one could say, learners had not enough information to deduce any requirement for expressivity). Overall then, any notion of (genetic) biases can only be considered in a situated context. Not only is the cultural expression of such biases dependent on the interactions between agents, but the (inanimate) environment itself is a factor as well.

As we have seen, the human subject experiments (Kirby et al. 2008; Perfors & Navarro 2011; Perfors & Navarro 2014), as well as their computational analogues discussed in Sections 10.2.1 and 10.2.2, consider cultural evolution of language by generating some form of utterance to express multiple meanings. When this requirement for expressivity is relaxed, the language is probable to converge to a non-compositional, underspecified state, potentially undermining the predictive power of the IL paradigm. A second point of attention on human subject IL studies is that utterances are defined to be discrete, i.e., decomposable in principle (e.g., a string is composed out of discrete characters). This decomposability might be less inherent to continuous speech. Finally, one could argue that the results rely on an existing predisposition for language processing in *adults*, as was already mentioned in Section 10.2.1. In reality, it is first language acquisition in infants that we should consider to be essential in the replication phase in cultural evolution language.

Results by Verhoef and de Boer (2011) and Verhoef, de Boer, and Kirby (2012) argue that these critiques are refutable by proposing a generalization of the traditional IL setup. Instead of having subjects tasked with conveying meanings using discrete utterances, they were required to replicate sounds with a slide whistle. This way, any implicit linguistic assumptions the subjects might impose on the task were attempted to be negated, while also lessening any inherent compositional structure by using a continuous signal. Following this approach, four chains of ten subjects were investigated, each showing a gradual increase in systematic recombination of signals and their utilization in forming compositional utterances. This was also confirmed by showing that the

normalized distance between subjects' received and produced utterances decreased the more often it was transmitted. Moreover, the Shannon entropy (i.e., the "information density" or "signal uncertainty"; Shannon 1948) of produced utterances decreased likewise, confirming the increasingly compositional nature of signals and their usage of shared elements.

The human subject experiments discussed so far seem to largely corroborate the predictions of the computational IL models. However, we argue that there are notable interpretational differences. First of all, in the computational models, agents had perfect memory. When they heard an utterance they were able to, in principle, replicate it flawlessly. Thus, the information bottleneck exerting selective pressure in cultural evolution is usually regarded as a logistic one: An agent can only transmit as many sounds as the bottleneck provides an opportunity for. However, the experiments with human subjects strongly suggest that the bottleneck is a result of limited memory capacity. Subjects are not able to learn all utterances by heart, even if they were exposed to all of them (Verhoef & de Boer 2011; Verhoef et al. 2012), and they are therefore forced to resort to making generalizations. Insofar, this bias in human subject IL seems to be, for a large part, cognitive in origin.

Intriguingly, although animal communication is often judged to be qualitatively different from human language (Hockett 1960), cultural evolution has also been established in nonhuman animals. A study by Fehér, Wang, Saar, Mitra, and Tchernichovski (2009) investigated "language" convergence in zebra finches, *Taeniopygia guttata*, using an experimental design largely similar to the IL studies. When growing up in isolation, the structure of the songs zebra finches produce is markedly different from those produced by zebra finches interacting with each other, demonstrating unusually long syllable duration, stuttering, and more broadband noise. In a first experiment, four zebra finches grown up in such a socially isolated situation served as tutors for juveniles. The juveniles, however, did not copy the tutors' song structure with high fidelity. Instead, the songs appeared more similar to the wild-type songs. When transmitted in succession from bird to bird, like in the IL experiments, song structure approached the wild type asymptotically. A similar convergence was found when isolates were founding, genetically as well as culturally, small colonies of zebra finches that allowed for horizontal transmission.

Conceptually, the zebra finch study directly addresses the influence of (genetic) biasing in cultural evolution. More specifically, there seems to be an intrinsic bias that forces the zebra finches to converge on the wild-type song structures, even when founded by isolates. This has a number of possible explanations. First, we suggest that the song structures in isolates and the tutored birds is resultant of one and the same bias and that the expression of this bias accumulates the more often it is transmitted (as would be expected from the amplification of weak biases following Bayesian IL using maximizers;

Section 10.2.2). In that case, however, the precise nature of the bias (e.g., anatomical, perceptual) is hard to pin down on face value. Alternatively, however, one might propose that the bias does not manifest in the isolates at all, implying some transmission factor between individuals is responsible for the wild-type convergence. For instance, it might be possible that the zebra finch mimicking behavior introduces a bias that is unused and therefore not expressed when birds develop in isolation. Perceptual biases, again likely not of importance in isolates' song structure, might introduce convergence to wild-type songs equally well. However, production biases (e.g., anatomical) seem less likely, since song production is a factor that is of comparable utilization in both isolates' as tutored birds' songs. Finally, there might be some sort of sexual selection in the colony model, e.g., females are more likely to mate with males that produce wild-type-like songs (although this seems unlikely given the observation that both (single-sex) chain as well as (mixed-sex) colony experiments produce similar (i.e., wild-type songs)).

10.2.4 *Self-Organization of Vowel Space*

The IL model (Sections 10.2.1 and 10.2.2) is a powerful approach and its predictions have been supported by human subject as well as animal models (Section 10.2.3). Furthermore, Bayesian IL (Section 10.2.2) allows the precise specification of priors that can be interpreted as representing innate biases with an (ultimately, at least partial) genetic basis. A problem however, as already addressed in Section 10.2.2, is its contradictory, strongly reductionist nature while describing language as an emergent property in a dynamical system. Moreover, the provided precision in defining priors constitutes a double-edged sword. More specifically, when modelling a natural system, the bias is often not known in detail and therefore hard to formalize in a Bayesian way. An alternative strategy is to look at cultural evolution of language from a *self-organizing*, dynamical systems perspective, in many ways resembling exemplar-based phoneme perception models (e.g., Kruschke 1992; Johnson 2005).

A study by de Boer (2000a, 2000b) showed that a self-organizing, population-level system can explain vowel dispersion in human languages accurately. Pairs of agents are iteratively selected to play an "imitation game." One agent initiates the game by transmitting a vowel that the imitator tries to match to one of its internally stored vowel prototypes using the Euclidean distance between formants. The imitator in turn produces a vowel based on the selected prototype, which the initiator in turn matches to a prototype as well. If both agents classify the same prototype, the communication game is a success and the imitator's prototype is shifted toward the perceived vowel. New prototypes are produced following unsuccessful games when the closest match

has a matching history too successful to be discarded. The prototype space is periodically cleansed based on the success history of prototypes.

The results show that, after 200 games, clusters in the vowel space start to form, explained by the fact that agents try to imitate each other while there is (cultural) selective pressure to have a maximum number of maximally distinguishable vowels. Thus, after 2,000 iterations, the vowel space is occupied by a few, tight clusters that are (between-cluster) maximally dispersed. An experiment where agents were periodically removed and new agents were inserted (where younger agents were able to change their vowel repertoire more easily, modeling first language acquisition in infants) showed similar results. Promisingly, the vowel clusters to which the experiments converged showed striking similarities with those seen in human languages (cf. Schwartz, Boë, Vallée & Abry 1997). Later studies used a similar self-organization approach, now including a temporal dimension (Zuidema & de Boer 2009; de Boer & Zuidema 2010). In doing so, it was shown that common sounds were reused but at different points in time. In effect, this strategy maximizes dispersion by using the temporal dimension as a distinguishability axis, leading to the development of (a system analogous to) combinatorial phonology that is often cited as unique to human language (Hockett 1960). Thus again, it appears some language properties can be explained without any need for nativism or biological evolution.

Although designed as a self-organizing system, the model by de Boer (2000a, 2000b) still incorporates an explicit, procedural definition of the language games agents play, both in terms of how agents communicate with each other as well as how they organize their phonemes internally. Moreover, the study therefore relies on the assumption that agents had the a priori capacity to communicate and that they experienced an explicit pressure to exploit this capacity. Thus, similarly to the IL models described earlier (Section 10.2.1), the studies do not explain glossogenesis, but are confined to the domain of cultural evolution of language when the capacity for it is already established. To address this, Oudeyer (2005a, 2005b) used an approach similar to de Boer (2000a, 2000b) but now basing agents' internals on self-organizing maps (SOMs) (Kohonen 1982, 2001). SOMs are often used for the purpose of dimensional reduction (akin to multidimensional scaling) and they are therefore applicable to map a high-dimensional input vector (e.g., the first few formant frequencies of a waveform) to some internal representation of fewer dimensions, argued to be biologically realistic. Subsequently, they can also be used to map an internal, high-dimensional representation onto a lower-dimensional output vector. More technically, an SOM consists of a layer of parallel neurons that each express a sensitivity for a particular input vector, where activation in one neuron will "bleed over" to neighboring ones (e.g., following a Gaussian distribution). Critically, these maps, as the name suggests, self-organize in that those neurons that yield the highest activation to an input vector tune themselves to respond to that

vector even more strongly. Following this architecture, Oudeyer (2005b) uses two of these SOMs, one perceptual and one motor map, in series to transform speech sounds into articulatory gestures. For this purpose, the two maps are fully connected and these connections are updated using a Hebbian learning rule (Hebb 1949). Pairs of agents are randomly selected: one produces a sound that the other one hears. Furthermore, the study used quasi-realistic¹⁰ vowel production and perception models, similar to those used in de Boer (2000a, 2000b).

In a first experiment (Oudeyer 2005a), however, vowel production was simplified, i.e., the mapping from the articulatory gestures to acoustics was linear. This nevertheless demonstrated that agents interacting with each other self-organize their vowel space around a few attractors, what could be called analogues to the prototypes in de Boer (2000a, 2000b). Interestingly however, this organization also emerged when agents were only allowed to self-talk, as in the babbling of infants. Expectedly, agents did not converge on a shared vowel dispersion pattern in this scenario, as was the case when agents were interacting with each other.

The experiment, while showing clear vowel convergence, did not reflect a vowel distribution seen in actual human populations. The second experiment (Oudeyer 2005a), however, using the more realistic gesture-to-vowel mapping, illustrated a more convincing picture in that respect: The distribution of vowel frequencies (i.e., how many vowel “phonemes” emerged in the simulations) clearly approximated that seen in human populations (Ladefoged & Maddieson 1998). Moreover, the dispersion in the vowel space showed a close resemblance to human language as well. Furthermore, in the follow-up (Oudeyer 2005b), combinatoriality (i.e., the reuse of segments at different points in time) in the vowel system was demonstrated by extending the model with a temporal neural map that self-organized through pruning the less active neurons (simulating, in a way, apoptosis – programmed cell death), allowing the production of (small and indeed combinatorial) sequences of vowels.

While the convergence on a discrete, one could claim phonological, vowel system occurs without the explicit intention to communicate, and even without any contact with peers whatsoever, the degree of realism of the “vocal tract” (linear versus parameterized) leads to different outcomes of the self-organizing process, given a tentative example of anatomical biasing. This issue was further

¹⁰ We use the term “quasi,” since the suggested realism is derived from numerical transformations. More precisely, vowels are produced using three parameters: lip rounding, tongue height, and tongue fronting. Vowel perception is based on a modified Barks transformation (Zwicker 1961) that accounts for relative narrow-band, high-frequency perceptual indiscriminability in human subjects. Thus, there is no simulated “physicality” in these models, but they remain relatively abstract, quasi-realistic instead.

emphasized in Oudeyer (2005a) that introduced a (“metabolic”) energy cost on vocal tract “displacements,” again resulting in different vowel dispersion patterns. None of these dispersion patterns were realistic, however. This, the author concludes, could have been expected, since dynamical systems are notoriously sensitive to initial conditions and perturbations of any kind. Failing to accurately model even a single biasing factor might therefore result in totally different outcomes. This realization then raises questions on the self-organization approach, particularly on what level of abstraction would be appropriate when modeling human language, i.e., with the aim to make testable predictions. More concretely, how would one estimate whether all relevant (genetic, anatomical, or other) biasing factors are accounted for while keeping the model, following Occam’s razor, as simple as possible? These questions remain as of yet unanswered.

Overall, and more theoretically, the self-organizing models describe a form of ontogenetic development – *phenotypic plasticity* – that can be conceptualized to work in conjunction with natural selection. It is not hard to imagine how the search space of human language features and parameters is enormous, and that natural selection in isolation is possibly not powerful enough to explore this space efficiently (Ball 1999). In that regard, ontogenetics such as self-organizing processes might actually amplify natural selection. For example, the studies by de Boer (2000a, 2000b) showed that language-like self-organization processes took place for a large range of parameters that determined the behavior of the SOMs. Thus, the target volume natural selection has to find to let language emerge is drastically inflated by application of this “local search.” In more ethological terms, this can be visualized by imaging natural selection traversing a search space of phenotypes in a relatively slow but robust fashion (i.e., emphasizing specialism), while ontogenetics allows for lifetime adaptations (i.e., emphasizing generalism) (Turney 1996). The realization that phenotypic plasticity might play a major role in biological evolution will probably have a massive effect on our understanding of evolutionary theory (Pigliucci 2007). For instance, this interaction might lead to another example of a situation where genes are being shielded from natural selection (since ontogenetics allow for a wider range of genetic polymorphisms to be effective). These interactions between phylogenetics, glossogenetics, and ontogenetics is discussed further in Section 10.3.

10.3 Cultural Feedback

So far, we have mainly discussed how genetics might bias cultural evolution of language, acting through neurocognition and the anatomy and physiology of the production and perception systems. This is a simplification of the natural

situation, however. For example, not only does human physiology shape language on a cultural level, but certain language features might conceivably become *assimilated* into the genome, and this genome, in turn, shapes language acquisition, processing, production, and perception. This thus effectively describes a dynamical system that would consist of three layers coupled in a feedback loop: genes, physiology, and culture.

The Baldwin Effect (Baldwin 1896) has generated considerable interest, describing the interaction between genes and phenotypic plasticity. First, the effect postulates that organisms might evolve to a state predisposed to adaptability. For example, organisms might develop complex nervous systems that allow them to cope with a dynamic environment. Secondly, the effect proposes that such ontogenetically acquired traits can be internalized into the organism's genome. Thus, even if we observe language modules or other biasing candidates in human anatomy and physiology, we cannot conclude a causal role. On the contrary, it might very well be the case that culturally expressed language features have caused the development of these (physiological) traits, instead of the other way around.

Investigations on the Baldwin Effect have a history of computational modelling. Hinton and Nowlan (1987), for instance, show that the use of phenotypic plasticity enables an organism to explore a search space of phenotypes and how, as mentioned in Section 10.2.3, this can be considered a form of local search on top of natural selection. Once optima are ontogenetically found, they increase an organism's fitness, after which they can subsequently be internalized into the genome, trading flexibility (generality) for optimality (specificity) when the situation requires. However, more recent experiments have cast doubts on the assimilation of language features expressed on a cultural level. For example, there are strong suggestions that in certain situations, described in Sections 10.2.3 and 10.2.4, ontogenetics or particular convergence patterns following cultural evolution might lead genes to be shielded from natural selection/assimilation.

In the domain of language evolution, the Baldwin Effect is often regarded from the *niche construction* interpretation, which states that the expression of stable language features (a stable cultural niche) exerts selective pressure on individuals to assimilate those features (Deacon 1997; Odling-Smee, Laland & Feldman 2003). Eventually, the predispositions to express particular language features might become so strong they appear "innate" (i.e., they are developed before birth, but the concept of innateness is notoriously complex; Mameli & Bateson 2006), via a process known as canalization (Waddington 1942). Such mechanisms might explain many cases of gene-culture coevolution such as the persistence of the adaptation to digest fresh milk in populations with dairy traditions (Richerson & Boyd 2008; Laland, Odling-Smee & Myles 2010; Richerson, Boyd & Henrich 2010; Richerson & Christiansen 2013).

The Baldwin Effect is conceptually closely related to the evolutionary mechanisms of adaptation and exaptation (Gould & Vrba 1982). To reiterate, adaptation describes the evolution of a novel, domain-specific trait (e.g., teeth for mastication), while exaptation relates to the co-option of existing traits for purposes they were not originally adapted for. An example of exaptation is the co-opting of feathers that some theropod dinosaurs used for thermo-regulation but with slight modifications were utilized to aid in flight in, e.g., birds (Ostrom 1976). Another example of exaptation might be the use of the lungs to produce vocalizations in animal communication or, more speculatively, of domain-general pattern recognition capabilities of the human brain in language cognition (Christiansen & Chater 2008). This latter example is particularly relevant because it is largely in line with the IL models discussed in Section 10.2, which describe how agents endowed with only general learning algorithms are able to produce complex languages without requiring any language modules.

As mentioned earlier in this section, some degree of modularity of mind might still arise out of selective pressure from sufficiently stable linguistic niches, leading to a possible genetic encoding to express the corresponding linguistic features. The required stability of this assimilation was investigated by Baronchelli, Chater, Christiansen, and Pastor-Satorras (2013), who used a particle model to run simulations on populations of “generalist” and “specialist” individuals. They showed that specialist individuals only evolved when they were confronted with a situation that allowed for little genetic and environmental change. When environmental change was larger, generalist individuals were favored. Drawing a parallel to language-gene coevolution, because language changes fast, this might imply that generalist speakers are favored, i.e., those that do not evolve language-specific adaptations.

Preluding the more general, abstract findings by Baronchelli et al. (2013), Chater, Reali, and Christiansen (2009) demonstrated that the assimilation of language features indeed requires a low rate of linguistic change in a computer simulation where language features exert selective pressure on the genome. More specifically, in the (unrealistic) scenario where the rate of linguistic change was equal to the rate of genetic change, assimilation is already substantially reduced when compared to when language was completely stable (let alone if the rate of linguistic change is higher). When the language was in turn partially (i.e., 50%) genetically determined, this provided a stabilizing influence that increased the assimilation rate. However, this ratio of genetic determinism appeared to be so (again, unrealistically) high that no distinction could be made between having a selective pressure from language and having no selective pressure at all (i.e., random genetic drift). In other words, the 50% genetic determinism scenario describes a situation in which the language can, for all intents and purposes, be said to be complete genetically determined, a situation that is extremely unlikely. Furthermore, high linguistic change favoured

the evolution of “neutral” alleles that allowed for more flexibility in expressing language features, while slow linguistic change lead to the evolution of alleles expressing such language features by genetic determinism (Baronchelli, Chater, Pastor-Satorras & Christiansen 2012). A similar finding was found when multiple populations were simulated that had interlingual contact: When an individual’s fitness was codetermined by its ability to learn a foreign population’s language, this again lead to the evolution of neutral alleles. Finally, the study also showed that features assimilated during a phase of slow linguistic change (e.g., during a proto-language) mutate into neutral alleles when the rate of change increases. Thus, the authors conclude that it is unlikely that assimilated remnants of a proto-language we might have spoken in the past still reside in our genome.

While interesting, most models on the Baldwin Effect make a number of simplifying assumptions. First of all, it is assumed that genes isomorphically correspond with linguistic features and that all these features had equal weight in expressing meaning. However, given what we know about the genetic bases of language (Fisher 2006), this assumption is almost certainly false. A second simplification is that all cultural linguistic features are equally stable. However, this has been long demonstrated not to be the case for human language (e.g., Maddieson 1984; Schwartz et al. 1997; Dunn et al. 2011), having lead to the suggestion of “universals” in the first place. Stable features such as these are of course more likely candidates for assimilation than unstable ones or those the stability of which is homogeneous and indistinct (such as in Chater et al. 2009, where this was recognized but not further addressed). Finally, the studies assume a simple, linear quantification of the cost of flexibility, i.e., the more plastic an organism, the longer it will take to arrive at the right phenotype. However, this quantifies only one of a number of cost associations with learning (Mayley 1996); none of them likely to have a linear signature.

To summarize, although the validity of Baldwinian niche construction with respect to cultural evolution of language is still debated, with multiple authors contradicting each other, we should consider that apparent domain-specific language modules or other adaptations, even if they were conclusively demonstrated, still do not imply the biological evolution of language-specific adaptations. Following this reasoning, if one were to discover apparent biasing factors, it might be the case that these only evolved after some stable linguistic feature became expressed, i.e., conforming to the selective pressure exerted from language itself instead of the other way around.

10.4 Conclusion

Biological and cultural evolution show striking similarities (Section 10.1.2). While not exact homologues, the principles of replication, variation, and

selection are found in both domains. This has the potential to explain linguistic features without the need to invoke “universals,” “language modules,” or the (purely) biological evolution of language. However, we are by no means stating that human languages do not, to some extent, show a degree of semi-universality, that certain cortical areas are more involved in language processing than others, or that biological natural selection is of no importance to explain the complexities of language. Nevertheless, the concept of cultural evolution is a large explanatory factor in itself and should be considered to interact in close conjunction with biological and physiological mechanisms. This insight is a shift away from the dominant, cognitivist perspective of speech perception and production, instead considering human language to (weakly) emerge from the interactions between speakers. Importantly, none of these speakers have any *a priori*, explicit internalization or predisposition to express linguistic features, but are exerting a social force on language convergence, shaping it in subtle, nonspecific ways. Extra-linguistic factors as well as innate biases speakers might have are therefore of importance, but we cannot hope to fully understand this convergence when taking a full reductionist approach in isolation.

Two frameworks are often used in computational modeling of cultural evolution. Iterated learning provides the benefit that it makes accurate, strong predictions. The “standard” IL experiments have, for instance, demonstrated the emergence of compositionality, recursion and the appearance of irregulars in Zipfian distributions (Section 10.2.1). Furthermore, using Bayesian agents in an IL framework provides a precise, tractable specification of biasing factors and, in doing so, illustrates how certain learning strategies might result in an amplification of weak biases, eventually hypothesized to lead to selective neutrality (or “shielding”) of bias strength (Section 10.2.2). Furthermore, the results are supported by trials on human subjects and by animal models (Section 10.2.3). However, as all models, IL forms an abstraction from the real world and rests on numerous assumptions. Relaxing those expectedly weakens some of the aforementioned predictions.

An alternative to using IL is to take the complex systems route, emphasizing self-organization to an even greater extent than IL and claiming that, for example, vowel space dispersion can be explained from the interactions between agents that might not even have any *a priori* conception of language, emphasizing glossogenesis that IL does not address (Section 10.2.4). More conceptually, these models illustrate how developmental, self-organizing processes can amplify the ability for natural selection to explore the vast search space of phenotypes, adding a degree of flexibility that selection alone does not provide. However, they also explicitly show that, an observation also relevant for IL, these systems are very sensitive to initial conditions and on-the-fly perturbations. In the end, our aim is to obtain a model that is as simple as possible,

without losing predictive power. This observed sensitivity of complex systems then questions what the appropriate level of abstraction for modeling cultural evolution of language is.

Extending this question from the topic of genetic biasing to gene-language coevolution raises more, similar issues (Section 10.3). It has long been recognized that ontogenetically acquired traits can be internalized into the genome. Language features have often been proposed to be candidates for this assimilation following Baldwinian niche construction, where plastic flexibility is traded for genetic rigidity when the environment is sufficiently stable. Computer models, however, question the feasibility of this hypothesis, showing that language features have to be very stable indeed or they need to be, for an unreasonable part, genetically determined for assimilation to occur. Once more, however, these models make a number of nontrivial assumptions, casting doubt on the validity of their level of abstraction. Overall, they, together with the divided literature from the biological sciences, illustrate that assimilation cannot be ruled out and therefore that, even if we would establish something like a cortical language module encoding word order, it would not imply its biological evolution as a causal factor. Instead, it might be that a (stable) cultural feature assumes that role and that any apparent adaptations are, upon close inspection, mere exaptations of existing traits.

To conclude this chapter, we hope to have shown the potential for using computer modeling to investigate the evolution of language as multiple, interacting domains subject to similar Darwinian principles, emphasizing the role of cultural evolution and the biasing effects genetics might have. While many questions remain to be answered, we regard this approach to be essential to fully account for the vast richness of human language and encourage a further transcendence of traditional disciplinary boundaries in this endeavour.

Acknowledgments

This chapter was funded by VIDI grant 276-70-022 of the Netherlands Organisation for Scientific Research (NWO) as part of the “Genetic Biases in Languages and Speech” research project. The authors are supported by the Max Planck Society.

References

- Baldwin, J. M. (1896), ‘A new factor in evolution’, *American naturalist* pp. 536–553.
Ball, P. (1999), *The self-made tapestry: Pattern formation in nature*, Oxford University Press.
Baronchelli, A., Chater, N., Christiansen, M. H. & Pastor-Satorras, R. (2013), ‘Evolution in a changing environment’, *PloS one* **8**(1), e52742.

- Baronchelli, A., Chater, N., Pastor-Satorras, R. & Christiansen, M. H. (2012), 'The biological origin of linguistic diversity', *PloS one* **7**(10), e48029.
- Bernardo, J. M. & Smith, A. F. (2009), *Bayesian theory*, Vol. 405, John Wiley & Sons.
- Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. & Atkinson, Q. D. (2012), 'Mapping the origins and expansion of the Indo-European language family', *Science* **337**(6097), 957–960.
- Burkett, D. and Griffiths, T. L. (2010), 'Iterated learning of multiple languages from multiple teachers', *The evolution of language: Proceedings of Evolang* pp. 58–65.
- Butcher, A. (2006), Australian Aboriginal languages: Consonant-salient phonologies and the 'place-of-articulation imperative', New York and Hove: Psychology Press, pp. 187–210.
- Calvin, W. H. (2002), *A brain for all seasons: Human evolution and abrupt climate change*, University of Chicago Press.
- Campbell, L. & Poser, W. J. (2008), *Language classification: History and method*, Cambridge University Press.
- Carroll, S. B. (2005), *Endless forms most beautiful: The new science of evo devo and the making of the animal kingdom*, number 54, W.W. Norton & Company.
- Chater, N., Reali, F. & Christiansen, M. H. (2009), 'Restrictions on biological adaptation in language evolution', *Proceedings of the National Academy of Sciences* **106**(4), 1015–1020.
- Chomsky, N. (1965), *Aspects of the theory of syntax*, number 11, MIT press.
- Chomsky, N. (1986), *Knowledge of language: Its nature, origin, and use*, Greenwood Publishing Group.
- Christiansen, M. H. & Chater, N. (2008), 'Language as shaped by the brain', *Behavioral and Brain Sciences* **31**(05), 489–509.
- Croft, W. (2000), *Explaining language change: An evolutionary approach*, Pearson Education.
- Dávid-Barrett, T. & Dunbar, R. (2013), 'Processing power limits social group size: Computational evidence for the cognitive costs of sociality', *Proceedings of the Royal Society B: Biological Sciences* **280**(1765).
- Dawkins, R. (1976), *The selfish gene*, Oxford University Press.
- de Boer, B. (2000a), 'Emergence of vowel systems through self-organisation', *AI Communications* **13**(1), 27–39.
- de Boer, B. (2000b), 'Self-organization in vowel systems', *Journal of Phonetics* **28**(4), 441–465.
- de Boer, B. & Fitch, W. T. (2010), 'Computer models of vocal tract evolution: An overview and critique', *Adaptive Behavior* **18**(1), 36–47.
- de Boer, B. & Zuidema, W. (2010), 'Multi-agent simulations of the evolution of combinatorial phonology', *Adaptive Behavior* **18**(2), 141–154.
- Deacon, T. (1997), *The symbolic species: The co-evolution of language and the brain*, number 202, WW Norton & Company.
- Dediu, D. (2008), 'The role of genetic biases in shaping the correlations between languages and genes', *Journal of Theoretical Biology* **254**(2), 400–407.
- Dediu, D. (2009), 'Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise?', *Journal of Theoretical Biology* **259**(3), 552–561.

- Dediu, D. (2011), 'Are languages really independent from genes? If not, what would a genetic bias affecting language diversity look like?', *Human Biology* **83**(2), 279–296.
- Dediu, D. & Levinson, S. C. (2013), 'On the antiquity of language: The reinterpretation of Neandertal linguistic capacities and its consequences', *Frontiers in Psychology* **4**.
- Dunn, M., Greenhill, S. J., Levinson, S. C. & Gray, R. D. (2011), 'Evolved structure of language shows lineage-specific trends in word-order universals', *Nature* **473**(7345), 79–82.
- Everett, C., Blasi, D. E. & Roberts, S. G. (2015), 'Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots', *Proceedings of the National Academy of Sciences* **112**(5), 1322–1327.
- Farrell, S., Wagenmakers, E.-J. & Ratcliff, R. (2006), '1/f noise in human cognition: Is it ubiquitous, and what does it mean?', *Psychonomic Bulletin & Review* **13**(4), 737–741.
- Fehér, O., Wang, H., Saar, S., Mitra, P. P. and Tchernichovski, O. (2009), 'De novo establishment of wild-type song culture in the zebra finch', *Nature* **459**(7246), 564–568.
- Ferdinand, V. and Zuidema, W. (2009), Thomas' theorem meets Bayes' rule: A model of the iterated learning of language, in 'Proceedings of the 31st Annual Conference of the Cognitive Science Society', Cognitive Science Society Austin, TX, pp. 1786–1791.
- Fisher, S. E. (2006), 'Tangled webs: Tracing the connections between genes and cognition', *Cognition* **101**(2), 270–297.
- Fodor, J. A. (1983), *The modularity of mind: An essay on faculty psychology*, MIT Press.
- Gould, S. J. & Vrba, E. S. (1982), 'Exaptation – a missing term in the science of form', *Paleobiology*, pp. 4–15.
- Gray, R. D. & Atkinson, Q. D. (2003), 'Language-tree divergence times support the Anatolian theory of Indo-European origin', *Nature* **426**(6965), 435–439.
- Griffiths, T. L. and Kalish, M. L. (2007), 'Language evolution by iterated learning with bayesian agents', *Cognitive Science* **31**(3), 441–480.
- Hamilton, W. D. (1963), 'The evolution of altruistic behavior', *American Naturalist* pp. 354–356.
- Hanke, D. (2004), 'Teleology: The explanation that bedevils biology', *Explanations: Styles of Explanation in Science*, pp. 143–155.
- Hebb, D. O. (1949), *The organization of behavior: A neuropsychological approach*, John Wiley & Sons.
- Henrich, J. & McElreath, R. (2003), 'The evolution of cultural evolution', *Evolutionary Anthropology: Issues, News, and Reviews* **12**(3), 123–135.
- Hinton, G. and Nowlan, S. (1987), 'How learning can guide evolution', *Complex Systems* **1**(1), 495–502.
- Hockett, C. (1960), 'The origin of speech', *Scientific American* **203**, 88–96.
- Johnson, K. (2005), Speaker normalization in speech perception, in *The handbook of speech perception*, John Wiley & Sons, pp. 363–389.
- Kirby, S., Cornish, H. & Smith, K. (2008), 'Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* **105**(31), 10681–10686.

- Kirby, S., Dowman, M. and Griffiths, T. L. (2007), 'Innateness and culture in the evolution of language', *Proceedings of the National Academy of Sciences* **104**(12), 5241–5245.
- Kirby, S. & Hurford, J. R. (2002), The emergence of linguistic structure: An overview of the iterated learning model, in *Simulating the evolution of language*, Springer, pp. 121–147.
- Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**(1), 59–69.
- Kohonen, T. (2001), *Self-organizing maps*, Vol. 30, Springer.
- Kröger, R. H. & Biehlmaier, O. (2009), 'Space-saving advantage of an inverted retina', *Vision Research* **49**(18), 2318–2321.
- Kruschke, J. K. (1992), 'Alcove: An exemplar-based connectionist model of category learning.', *Psychological Review* **99**(1), 22.
- Ladd, D. R., Dediu, D. & Kinsella, A. R. (2008), 'Languages and genes: Reflections on biolinguistics and the nature-nurture question', *Biolinguistics* **2**(1), 114–126.
- Ladefoged, P. (1984), Out of chaos comes order? Physical, biological, and structural patterns in phonetics, in A. Cohen & M. van den Broecke, eds, *Proceedings of the Tenth International Congress of Phonetic Sciences*, Foris Publications: Dordrecht, Holland, pp. 83–95.
- Ladefoged, P. & Maddieson, I. (1998), 'The sounds of the world's languages', *Language* **74**(2), 374–376.
- Laland, K. N., Odling-Smee, J. & Myles, S. (2010), 'How culture shaped the human genome: Bringing genetics and the human sciences together', *Nature Reviews Genetics* **11**(2), 137–148.
- Levinson, S. C. & Gray, R. D. (2012), 'Tools from evolutionary biology shed new light on the diversification of languages', *Trends in Cognitive Sciences* **16**(3), 167–173.
- Lin, C. & Shu, F. H. (1964), 'On the spiral structure of disk galaxies.', *The Astrophysical Journal* **140**, 646.
- Maddieson, I. (1984), *Patterns of sounds*, Cambridge University Press.
- Mameli, M. & Bateson, P. (2006), 'Innateness and the sciences', *Biology and Philosophy* **21**(2), 155–188.
- Mayley, G. (1996), The evolutionary cost of learning, in *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, pp. 458–467.
- Mesoudi, A. & Whiten, A. (2008), 'The multiple roles of cultural transmission experiments in understanding human cultural evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1509), 3489–3501.
- Miller, G. (2001), 'The mating mind: How sexual choice shaped the evolution of human nature', *Psycoloquy* **12**(8), 1–15.
- Odling-Smee, F. J., Laland, K. N. & Feldman, M. W. (2003), *Niche construction: The neglected process in evolution*, number 37, Princeton University Press.
- Okasha, S. (2006), *Evolution and the Levels of Selection*, Vol. 16, Clarendon Press Oxford.
- Ostrom, J. H. (1976), 'Archaeopteryx and the origin of birds', *Biological Journal of the Linnean Society* **8**(2), 91–182.
- Oudeyer, P.-Y. (2005a), 'The self-organization of combinatoriality and phonotactics in vocalization systems', *Connection Science* **17**(3–4), 325–341.

- Oudeyer, P.-Y. (2005b), 'The self-organization of speech sounds', *Journal of Theoretical Biology* **233**(3), 435–449.
- Pagel, M., Atkinson, Q. D. & Meade, A. (2007), 'Frequency of word-use predicts rates of lexical evolution throughout Indo-European history', *Nature* **449**(7163), 717–720.
- Perfors, A. (2012), 'Bayesian models of cognition: What's built in after all?', *Philosophy Compass* **7**(2), 127–138.
- Perfors, A. & Navarro, D. J. (2011), Language evolution is shaped by the structure of the world, in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Cognitive Science Society.
- Perfors, A. and Navarro, D. J. (2014), 'Language evolution can be shaped by the structure of the world', *Cognitive Science* **38**(4), 775–793.
- Pigliucci, M. (2007), 'Do we need an extended evolutionary synthesis?', *Evolution* **61**(12), 2743–2749.
- Pinker, S. & Bloom, P. (1990), 'Natural language and natural selection', *Behavioral and Brain Sciences* **13**(4), 707–727.
- Richerson, P. J. & Boyd, R. (2008), *Not by genes alone: How culture transformed human evolution*, University of Chicago Press.
- Richerson, P. J., Boyd, R. & Henrich, J. (2010), 'Gene-culture coevolution in the age of genomics', *Proceedings of the National Academy of Sciences* **107**(Supplement 2), 8985–8992.
- Richerson, P. J. & Christiansen, M. H. (2013), *Cultural evolution: Society, technology, language, and religion*, MIT Press.
- Schwartz, J.-L., Boë, L.-J., Vallée, N. & Abry, C. (1997), 'Major trends in vowel system inventories', *Journal of Phonetics* **25**(3), 233–253.
- Shannon, C. E. (1948), *The mathematical theory of communication*, University of Illinois Press.
- Smith, K. (2001), The evolution of learning mechanisms supporting symbolic communication, in *CogSci2001, the 23rd Annual Conference of the Cognitive Science Society*, Citeseer.
- Smith, K. (2009), Iterated learning in populations of bayesian agents, in *Proceedings of the 31st annual conference of the cognitive science society*, Austin, TX: Cognitive Science Society, pp. 697–702.
- Smith, K. and Kirby, S. (2008), 'Cultural evolution: Implications for understanding the human language faculty and its evolution', *Philosophical Transactions of the Royal Society B: Biological Sciences* **363**(1509), 3591–3603.
- Smith, K., Tamariz, M. & Kirby, S. (2013), Linguistic structure is an evolutionary trade-off between simplicity and expressivity, in *Proceedings of Cogsci 2013*, pp. 1348–1353.
- Turney, P. (1996), Myths and legends of the Baldwin Effect, in *Proceedings of the Workshop on Evolutionary Computing and Machine Learning at the 13th International Conference on Machine Learning*, pp. 135–142.
- Verhoef, T. & de Boer, B. (2011), Cultural emergence of feature economy in an artificial whistled language, in *Proceedings of the 17th international congress of phonetic sciences. Hong Kong: City University of Hong Kong*, pp. 2066–2069.
- Verhoef, T., de Boer, B. & Kirby, S. (2012), Holistic or synthetic protolanguage: Evidence from iterated learning of whistled signals, in *The evolution of language:*

- Proceedings of the 9th international conference (EVOLANG9)*, World Scientific, pp. 368–375.
- Waddington, C. H. (1942), ‘Canalization of development and the inheritance of acquired characters’, *Nature* **150**(3811), 563–565.
- Williams, G. C. (1966), *Adaptation and natural selection: A critique of some current evolutionary thought*, Princeton University Press.
- Wilson, D. S. & Wilson, E. O. (2008), ‘Evolution “for the good of the group”’, *American Scientist* **96**(5), 380–389.
- Wynne-Edwards, V. C. (1962), *Animal dispersion in relation to social behaviour*, Hafner Pub. Co.
- Wynne-Edwards, V. C. (1986), *Evolution through group selection*, Blackwell Scientific.
- Zipf, G. K. (1949), *Human behavior and the principle of least effort*, Addison-Wesley.
- Zuidema, W. & de Boer, B. (2009), ‘The evolution of combinatorial phonology’, *Journal of Phonetics* **37**(2), 125–144.
- Zwicker, E. (1961), ‘Subdivision of the audible frequency range into critical bands (Fre quenzgruppen)’, *The Journal of the Acoustical Society of America* **33**(2), 248–248.

11 Transparency versus Processing Efficiency: A Case Study on German Declension

Remi van Trijp

Abstract

Ambiguity is one of the most fascinating mysteries of human language that divides the linguistic research field in roughly two camps: the mainstream view, which considers ambiguity to be undesirable for communication; and the cognitive-functional view, which argues that ambiguity allows for more efficient communication. This chapter subscribes to the cognitive-functional view and presents a case study on the German declension system, which is notorious for its ambiguity through the use of syncretic case forms. Through a methodology based on computational reconstruction, the paper suggests that the current declension system outperforms its historical predecessors in terms of efficient communication, while featuring a labor-saving distribution of morphological marking across articles, adjectives, and nouns.

11.1 Introduction

Intuitively speaking, successful communication requires languages to exhibit a transparent mapping between meaning and form. Cross-linguistic research, however, shows that such transparency is the exception rather than the rule (Leufkens, 2015). Indeed, computational linguists will sometimes jokingly admit that they only have three problems to solve: ambiguity, ambiguity, and ambiguity.¹

Ambiguity is a major puzzle that divides the study of natural language in roughly two views (Winkler, 2015). One view, most outspokenly assumed by Chomskyan linguistics, considers ambiguity to be harmful for communication – an assumption that is sometimes used for arguing that language is in fact “poorly designed” for communication (Chomsky, 2008, p. 136). The opposing view, defended in cognitive-functional approaches to language, argues that

¹ Also see Manning and Schütze (1999, Chapter 1) on why ambiguity is so difficult for research on natural language processing.

ambiguity makes inferential communication systems more efficient (Piantadosi et al., 2012).

This study subscribes to the cognitive-functional approach and argues that ambiguity may lead to greater efficiency in language processing. As a case study, the chapter focuses on why *case syncretism* has emerged in the German declension system. Syncretism occurs when the same form can be mapped onto different grammatical functions. One example is the English suffix *-s*, which can be used as a number marker in nouns (e.g. *cats*) or as a conjugation marker in verbs *discovers*. Often enough, these functions may be contradictory: as a nominal marker, *-s* expresses plurality, but when used as a verbal suffix, it expresses agreement with a singular subject. The German declension system in particular is notorious for this kind of syncretism, which has puzzled many linguists for decades (see a.o. Bierwisch, 1967; Blevins, 1995; Wiese, 1996; Wunderlich, 1997; Müller, 2001; Daniels, 2001; Müller, 2002; Crysmann, 2005; Dalrymple et al., 2009).

This paper defends the hypothesis that syncretism emerged in the German declension system for efficiency reasons through a methodology that is based on computational reconstruction: it reconstructs a computational processing model of the Old High German declension system (Wright, 1906, OHG; 500–1100 AD) and of the contemporary German or New High German (NHG; from 1350 AD on), which can then be compared to each other in terms of communicative measures such as transparency, disambiguation power, processing efficiency, and so on. The remarkable result is that the OHG system, despite its greater degree of transparency, is less adapted for communicative efficiency than the current system is. The experiments reported in this essay develop on earlier findings (van Trijp, 2011b, 2013), but investigate more closely the importance of developing adequate representations and processing models for analyzing linguistic phenomena.

11.2 German Declension: Not as Awful as It Seems

In his 1880 essay, the American author Mark Twain famously complained that *The awful German language* is the most “slipshod and systemless, and so slippery and elusive to grasp” language of all. The German declension system, which combines the dimensions of case, number, and gender, seems to fit the bill, as it features a lot of syncretic forms. This can be seen in Table 11.1, which shows the German paradigm of definite articles. For instance, the definite article *der* can be used as a determiner for nouns that are (a) nominative-singular-masculine, (b) dative-singular-feminine, (c) genitive-singular-feminine, and (d) genitive-plural. The language only uses six different articles for a theoretical upper bound of 24 possible combinations (4 case distinctions \times 2 number distinctions \times 3 gender distinctions). Even when discarding gender distinctions

Table 11.1 *German definite articles are marked for case, number, and gender. Theoretically, the paradigm contains 24 different possible combinations (4 case distinctions × 2 number distinctions × 3 gender distinctions), but it is usually represented as containing 16 different cells because there are no gender distinctions for plural forms. Nevertheless, the language only uses six forms for filling these sixteen cells.*

Case	SG-M	SG-F	SG-N	PL
NOM	<i>der</i>	<i>die</i>	<i>das</i>	<i>die</i>
ACC	<i>den</i>	<i>die</i>	<i>das</i>	<i>die</i>
DAT	<i>dem</i>	<i>der</i>	<i>dem</i>	<i>den</i>
GEN	<i>des</i>	<i>der</i>	<i>des</i>	<i>der</i>

for plural nouns (leading to 16 distinctions), each form occupies at least two cells and their corresponding functions.

11.2.1 Some Observations about the Declension System

Let's first review some facts about the German declension system, more particularly how it applies to articles, adjectives, and nouns. German is predominantly dependent-marking, which means that most of the burden of case inflection is carried by articles and/or adjectives rather than by the head noun of an NP.

Nouns. German nouns are divided into three gender classes: masculine (e.g., *der Mann* “the man”), feminine (*die Frau* “the woman”), and neuter (*das Kind* “the child”). Plural forms are typically marked through a suffix (e.g., *die Frau-en* “the women”) and/or adding an umlaut (e.g., *die Mause* “the mouse” vs. *die Mäuse* “the mice”), although some nouns take zero marking (e.g., *der Löffel* “the spoon” vs. *die Löffel* “the spoons”). German plural marking is a complex problem in its own right (see, e.g., Daelemans and Van den Bosch, 2005, Chapter 3, for experiments on the acquisition of German plurals), but for our purposes it suffices to remember that nouns usually make a distinction between singular and plural forms.

Case marking is sparse, involving only a small number of highly syncretic forms. Most feminine nouns are unmarked for case (except dative plurals; see further later). Masculine and neuter nouns usually have genitive case endings (e.g., *des Kind-s* “the child's”), unless they are so-called *n-nouns*, which take the suffix *-(e)n* for all non-nominative cases (e.g., *der Junge* “the boy.NOM” vs. *der Junge-n* “the boy-ACC/DAT/GEN”). Finally, plural nouns take the suffix *-n*

Table 11.2 *German indefinite articles follow roughly the same declension pattern as the definite articles. Notable differences are the zero marking for nominative masculine and neuter forms and the accusative neuter form. The indefinite article also has no plural form, as is the case in English.*

Case	SG-M	SG-F	SG-N
NOM	<i>ein</i>	<i>eine</i>	<i>ein</i>
ACC	<i>einen</i>	<i>eine</i>	<i>ein</i>
DAT	<i>einem</i>	<i>einer</i>	<i>einem</i>
GEN	<i>eines</i>	<i>einer</i>	<i>eines</i>

in the dative case, unless they already end in *-n* or *-s* (e.g., *den Kind-er-n* “the child-PL-DAT”).

Indefinite Articles. The declension of the indefinite article *ein*, shown in Table 11.2, is in most cases similar to that of the definite article. The main differences are the lack of a plural form, and zero marking for nominative masculine and neuter, and accusative neuter forms.

Adjectives. German predicative adjectives are not declined for case. Attributive adjectives, on the other hand, follow three declension patterns depending on the composition of the NP in which they occur, as can be seen in Table 11.3. First, adjectives take strong declension if there is no preceding determiner. If, however, the adjective is preceded by a definite article (or other determiner with the same declension pattern), it takes weak declension. In this case, only two case suffixes are used: *-e* and *-en*. Finally, if the preceding determiner is indefinite, the adjective takes a mixed declension pattern.

A more general way of viewing the declension of German attributive adjectives is to say that the adjective takes on case-number-gender marking if such marking has not yet been taken by a preceding determiner. That means that if the determiner has no case suffix (e.g., *ein*), the adjective uses strong declension, as in *ein jung-er Mann* “a young-NOM.SG.M man.”

11.2.2 How to Handle Case Syncretism

Many scholars have worked on the problem of the German declension system through formal grammar approaches (Bierwisch, 1967; Blevins, 1995; Wiese, 1996), and the system’s syncretism has proven such a challenge (Ingria, 1990) that it has inspired a lot of innovations in the apparatus employed by formal grammars (Crysman, 2005; Daniels, 2001; Heinz and Matiasek, 1994; Müller, 1999, 2001; Sag, 2003). As is clear from the previous section, the challenge stems from the fact that each case form maps onto multiple, conflicting values that interweave three dimensions: case, number, and gender.

Table 11.3 *Attributive adjectives take three different declension patterns depending on the presence and kind of a preceding determiner.*

(a) Strong declension:				
Case	SG-M	SG-F	SG-N	PL
NOM	-er	-e	-es	-e
ACC	-en	-e	-es	-e
DAT	-em	-er	-em	-en
GEN	-en	-er	-en	-er
(b) Weak declension:				
Case	SG-M	SG-F	SG-N	PL
NOM	-e	-e	-e	-en
ACC	-en	-e	-e	-en
DAT	-en	-en	-en	-en
GEN	-en	-en	-en	-en
(c) Mixed declension:				
Case	SG-M	SG-F	SG-N	PL
NOM	-er	-e	-es	-en
ACC	-en	-e	-es	-en
DAT	-en	-en	-en	-en
GEN	-en	-en	-en	-en

Disjunctive Feature Representation. One traditional solution is to use *disjunctive feature representation*, in which multifunctionality is represented through listing the possibilities as disjunctions (i.e., separate alternatives). For example, the article *die* covers the nominative and accusative feminine singular case, or all plural nominative and accusative nouns. Example (1) shows a feature structure (adopted from Karttunen, 1984, p. 30) with feature-value pairs between square brackets. Disjunctions are presented by enclosing the alternatives in curly brackets ({}).

$$(1) \quad \begin{array}{l} \text{AGREEMENT} \\ \left[\begin{array}{l} \text{GENDER } f \\ \text{NUM } sg \\ \left\{ \begin{array}{l} \text{NUM } pl \\ \text{CASE } \{ nom \text{ acc } \} \end{array} \right. \end{array} \right] \end{array}$$

Despite its elegance, disjunctive feature representation is not without flaws. Crysmann (2005) argues that the grammarian is often forced to make arbitrary implementation decisions. Moreover, disjunctive features are considered

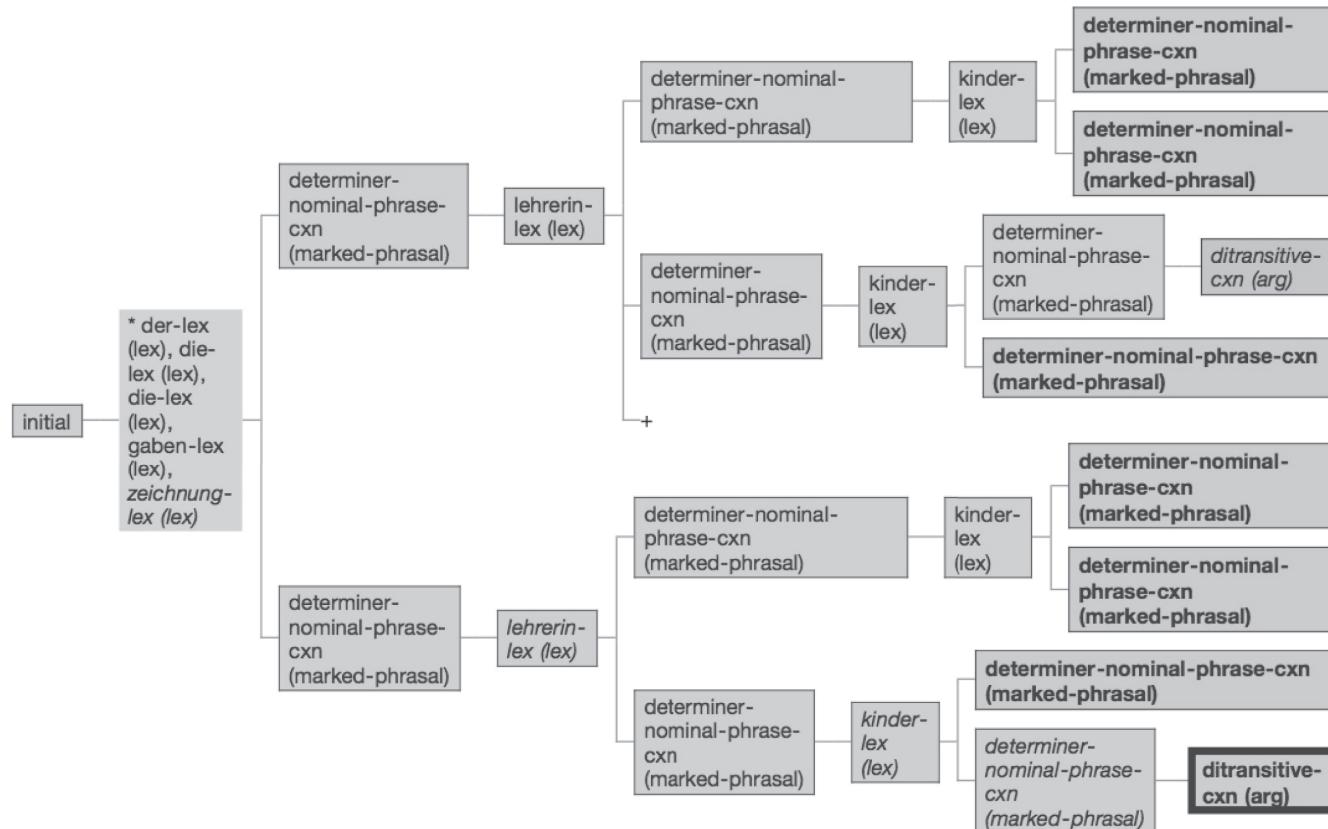


Figure 11.1 Parsing of the utterance *Die Kinder gaben der Lehrerin die Zeichnung* “The children gave the drawing to the (female) teacher.” As can be seen, disjunctions force splits in the search tree regardless of syntactic context.

to be computationally expensive representations (Flickinger, 2000). In fact, the computational complexity of unifying disjunctive features is NP-complete (Ramsay, 1990), which means grammar engineers can only reach efficient implementations by carefully managing the input and processing heuristics, or to resort to approximate results with faster algorithms (e.g., Carter, 1990; Ramsay, 1990). Some implementations have decided to eliminate disjunctions altogether whenever possible (Flickinger, 2000; Crysmann, 2005).

Figure 11.1 illustrates the problem. The figure shows the search tree for parsing the utterance *Die Kinder gaben der Lehrerin die Zeichnung* “the children gave the drawing to the (female) teacher.” The example uses a mini-grammar for German that consists of only six lexical entries: the definite articles *die* and *der*, the nouns *Kinder* “children,” *Lehrerin* “female teacher” and *Zeichnung* “drawing,” and the verb form *gaben* “gave.PL.” All lexical entries use disjunctive feature representation for their agreement features case, gender, and number similar to examples (1) mentioned earlier. Additionally, the grammar contains a Determiner-Noun construction that imposes agreement between the determiner and its head noun, and a ditransitive construction that captures the argument structure of the utterance.

What happens is that the disjunctions cause a split in the search tree whenever there are multiple alternatives possible. For example, *die Kinder* could be nominative or accusative plural, *der Frau* could be dative or genitive singular, and *die Zeichnung* could be nominative or accusative singular. This means that the search engine potentially has to consider seven false parses before the correct one is found. Additionally, every time a branch splits, the search space balloons accordingly because the search algorithm has to consider alternative orderings of applying the same constructions. The pluses in the figure stand for these alternative branches that lead to duplicate nodes in the search tree. Their full expansion is not shown because of space limitations, but it should be obvious that detecting and pruning such duplicate nodes is a costly matter in terms of processing effort.

These efficiency issues also suggest that this search process is implausible from a psycholinguistic point of view because the example utterance is unambiguous for German speakers: *die Kinder* is the only candidate for being the subject because it is the only noun phrase that agrees with the main verb. This leaves only the accusative slot open for *die Zeichnung*, and finally, *der Lehrerin* is unambiguously assigned dative case by the verb. In other words, the search tree does not reflect the processing choices that a natural language user would make as well, and they cause ambiguities even when the syntactic context is clear for native speakers.

Type Hierarchies. Type hierarchies have been proposed as a more elegant and efficient alternative to disjunctions in most contemporary

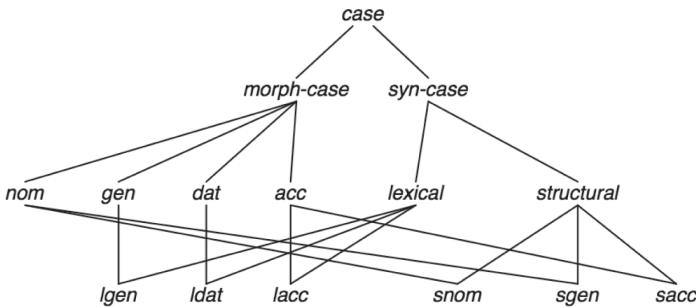


Figure 11.2 A type hierarchy proposed for German case agreement (Heinz and Matiasek 1994, figure adopted from Müller 2001).

grammar formalisms. These so-called *typed* feature structure grammars classify linguistic items in terms of types, which themselves are usually organized in a multiple inheritance network (see Figure 11.2). For each type, particular constraints (“type constraints”) can be defined, and each type has to satisfy the type constraints of all of its supertypes plus every constraint imposed on the type itself. A formalism’s type system thus “acts as the defining framework for the rest of the grammar. For instance, it determines which structures are mutually compatible and which features can occur, and it sets up an inheritance system which allows generalizations to be expressed” (Copestake, 2002, p. 35). Even though type hierarchies do not exclude the use of disjunctions, they have sometimes been presented as a way to eliminate disjunctions whenever possible because they significantly increase efficiency (Flickinger, 2000). For German as well, various type hierarchies have been proposed (Heinz and Matiasek, 1994; Daniels, 2001; Müller, 2001).

However, the grammar of German allows for constructions that are not easily captured through a type hierarchy. Example (2) shows a coordination construction between two noun phrases. Here, *dem Mann* and the pronominal *der* share the dative case, but they differ in gender, which is not possible using a single type hierarchy for the case-number-gender combination, because structure sharing in this approach forces types to agree in number and gender as well (Müller, 2001). Solutions vary from introducing additional features (*ibid.*) to positing relational constraints (Daniels, 2001), but all of them return at least partially to disjunctive feature representation and therefore neutralize the efficiency gain of type hierarchies (Crysman, 2005).

- (2) *Ich helfe der und dem Mann.*
 I help the.DAT.SG.F and the.DAT.SG.M man
 “I help this one and the man.”

A second problem for type hierarchies is feature indetermination, as illustrated in (3) (Pullum and Zwicky, 1986; quoted from Crysmann, 2005, p. 24):

- (3) *Er findet und hilft Frauen.*
 he finds.A and helps.D women.A/D
 “He finds and helps women.”

The verb *finden* “to find” normally takes an accusative complement, whereas *helfen* “to help” takes a dative complement. *Frauen* is underspecified and can be both accusative or dative. A sentence such as **Er findet und hilft Kindern* “He finds and helps children,” on the other hand, is ungrammatical because *Kindern* can only be dative and hence clashes with the requirements of the verb *finden*. Based on such examples, it has been argued by Ingria (1990) that unification is not the best technique for syntactic agreement and case assignment, and that compatibility checks are needed instead.

Researchers have gone to great lengths to counter Ingria’s claim, especially within the HPSG framework (Müller, 1999; Daniels, 2001; Sag, 2003). One solution is to augment the type hierarchy to explicitly contain neutral (or indeterminate) types (Levine et al., 2001) that can act as if they have multiple values. In example (3), the word *Frauen* would have a neutral feature so it may act as though it has both dative and accusative feature values.

Unfortunately, it is very hard to decide when types should be treated as neutral (i.e. indetermined) or ambiguous. Moreover, as argued by Crysmann (2005), such a solution leads to drastic increases in the amount of lexical ambiguity. Crysmann writes that the apparent incompatibility of feature indetermination and underspecification cannot be overcome using a single type hierarchy. Instead, he proposes two partially independent hierarchies to enable the Case values to be isolated from the Number-Gender values.

Distinctive Feature Matrices. In recent years, the puzzle of German syncretism has taken a promising turn through the use of *distinctive feature matrices* (Dalrymple et al., 2009; van Trijp, 2011b), which is an elaboration on the proposal by Ingria (1990) to no longer treat the value of a feature such as Case as atomic, but as an array of values. In this chapter, I adopt distinctive feature matrices as implemented in Fluid Construction Grammar (van Trijp, 2011b).

Returning to the example *Die Kinder gaben der Lehrerin die Zeichnung* “the children gave the drawing to the teacher” and ignoring genitive for the time being, the case feature of the definite article *die* and the noun *Zeichnung* could be represented as follows:

Table 11.4 *The feature matrix for German case.*

Case	SG-M	SG-F	SG-N	PL
?NOM	?nom-s-m	?nom-s-f	?nom-s-n	?nom-pl
?ACC	?acc-s-m	?acc-s-f	?acc-s-n	?acc-pl
?DAT	?dat-s-m	?dat-s-f	?dat-s-n	?dat-pl
?GEN	?gen-s-m	?gen-s-f	?gen-s-n	?gen-pl

- (4) die:
$$\text{CASE} \begin{bmatrix} \text{nom} & ?\text{nom} \\ \text{acc} & ?\text{acc} \\ \text{dat} & - \end{bmatrix}$$
- (5) Zeichnung:
$$\text{CASE} \begin{bmatrix} \text{nom} & ?\text{nom} \\ \text{acc} & ?\text{acc} \\ \text{dat} & ?\text{dat} \end{bmatrix}$$

The aforementioned representation, which is a simplification for illustration purposes only, captures the fact that *die* is ambiguous for nominative and accusative through the use of variables (i.e., symbols with a question mark), but that it excludes dative (through the symbol $-$). *Zeichnung* can be assigned any of these three cases.

Remember from Figure 11.1 that disjunctive feature representation forces a split in the search tree between a nominative and accusative reading of *die Zeichnung*, even though the syntactic context is unambiguous. Feature matrices avoid this problem because they make use of underspecification. Unifying *die* and *Zeichnung* leads to the following feature matrix, which can still be assigned nominative or accusative case later on, but which already excludes dative:

- (6) die Zeichnung:
$$\text{CASE} \begin{bmatrix} \text{nom} & ?\text{nom} \\ \text{acc} & ?\text{acc} \\ \text{dat} & - \end{bmatrix}$$

Let us now use this simple idea and apply it to the German declension paradigm. Each cell in the paradigm can be assigned a variable. This leads to the distinctive feature matrix for German as shown in Table 11.4. The first column of the table shows the dimension of case. All the other columns of the table feature cells that represent a specific case-gender-number combination. For example, the variable ?nom-s-m stands for “nominative singular masculine.” Since plural forms do not mark differences in gender, only one plural cell is included for each case.

Table 11.5 *The feature matrix for der.*

Case	S-M	S-F	S-N	PL
?nom-s-m	?nom-s-m	–	–	–
–	–	–	–	–
?dat-s-f	–	?dat-s-f	–	–
?gen	–	?gen-s-f	–	?gen-pl

Table 11.6 *The feature matrix for Lehrerin.*

Case	S-M	S-F	S-N	PL
?nom-s-f	–	?nom-s-f	–	–
?acc-s-f	–	?acc-s-f	–	–
?dat-s-f	–	?dat-s-f	–	–
?gen-s-f	–	?gen-s-f	–	–

Each linguistic item fills in as much information as possible in this case matrix. For example, the definite article *der* underspecifies its potential values and rules out all other options through “–,” as shown in Table 11.5. Note that the variable name for the nominative case ?nom-s-m is the same as the one for the cell of nominative-singular-masculine, which means that if the article unifies with a masculine noun, it is automatically disambiguated as a nominative article, and vice versa, if the article is assigned nominative case, we can infer that it is masculine. The same goes for the dative case.

The string *Lehrerin* “teacher.F.SG” rules out all plural forms but allows any case assignment. Since this noun is feminine, the single-dimension variables for case are the same ones as those that fill the singular-feminine cells in the matrix, as shown in Table 11.6.

Unification of *der* and *Lehrerin* only leaves the cells for dative and genitive feminine-singular open. In other words, *der Lehrerin* can only fill a dative or genitive slot. Other constructions may then later assign a “+” value to one of the two cases. The resulting feature matrix is shown in Table 11.7.

Table 11.7 *The feature matrix for der Lehrerin.*

Case	S-M	S-F	S-N	PL
–	–	–	–	–
–	–	–	–	–
?dat-s-f	–	?dat-s-f	–	–
?gen-s-f	–	?gen-s-f	–	–



Figure 11.3 The search tree for *Die Kinder gaben der Lehrerin die Zeichnung* using feature matrices in the grammar.

The efficiency of this technique is illustrated in Figure 11.3, which shows the search tree for parsing the same utterance *Die Kinder gaben der Lehrerin die Zeichnung* using feature matrices in the grammar. As opposed to the search with disjunctions (see Figure 11.1), feature matrices do not cause splits in the search tree unless there is an actual ambiguity in the language. Instead, they postpone commitment to any particular value as long as possible and thus allow information and constraints to be filled in by every part of the linguistic inventory.

Besides the enormous efficiency gain and a more plausible search process, feature matrices only require unification as the standard processing mechanism without additional sources for checking compatibility of information. Moreover, as demonstrated by van Trijp (2011b), the technique is also expressive enough to deal with those cases where traditional solutions are struggling.

Representing the value of the case feature as a complex array is not only an elegant solution that rules out spurious ambiguity caused by inadequate representations; it also fits with recent psycholinguistic evidence that suggests that human language processing is made more efficient by considering paradigmatic oppositions (Clahsen et al., 2001; Eisenbeiss et al., 2005/2006). In other words, the German declension system – which had been branded as partially accidental and non-systematic – turns out to be not as awful as it seems at first sight.

11.3 Evaluating the Efficiency of Syncretism

Now that we have established a formal representation for case syncretism that eliminates spurious ambiguity, we need to evaluate whether syncretic forms are more or less efficient than transparent forms. This can be achieved through a novel methodology based on *computational reconstruction*. The methodology is based on the following steps (van Trijp, 2013):

1. Construct a corpus that contains relevant data for the research topic (here: constructions in which case declension matters).
2. Operationalize relevant assessment criteria. This paper uses the criteria based on language usage cue reliability, disambiguation power, processing efficiency, auditory distinctiveness, and ease of pronunciation.
3. Reconstruct a grammar sample of the language under investigation, which is capable of parsing and producing the structures found in the corpus, and evaluate its performance using the assessment criteria.

4. Replace the part of the grammar under investigation with an alternative variation and evaluate its performance using the assessment criteria.
5. Compare the results.

11.3.1 Experimental Setup

In the current experiment, we are interested in the performance of the German declension system. More specifically, we wish to test whether the system is more or less efficient than a comparable declension system that features less syncretism and more transparency in its meaning-form mappings.

The Corpus. The most important function of case marking is to identify “who does what to whom,” that is, to mark the relations between a predicate and its arguments. The experiment therefore involves a corpus of declarative German utterances. There are three basic utterance *types*:

1. Ditransitive: NOM – Verb – DAT – ACC
e.g., Die Kinder gaben der Frau die Zeichnung. (“The children gave the drawing to the woman.”)
2. Transitive (a): NOM – Verb – ACC
e.g., Die Frau sah den Mann. (“The woman saw the man.”)
3. Transitive (b): NOM – Verb – DAT
e.g., Der Mann hilft der Frau. (“The man helps the woman.”)

The meanings that the agents need to express in production consist of a verb (e.g., to help), its “participant roles” (e.g., a helper-role and a beneficiary-role) and its arguments (e.g., a man and a woman). Meanings are represented using a first-order predicate logic (Steels, 2004; van Trijp, 2011a):

$$(7) \quad \exists ev, x, y: \text{help}(ev), \text{helper}(ev, x), \text{beneficiary}(ev, y), \text{man}(x), \text{woman}(y)$$

All nouns in the corpus have a distinct lexical form for singular and plural (e.g., *Mann* vs. *Männer*; “man” vs. “men”), but are considered to be unmarked for case. That is, even if they did carry case information, this information would be ignored in the experiments. Each utterance type features subtypes that involve a unique type combination for the dimensions of number and gender, which yields 216 unique utterance subtypes for the ditransitive as follows:

	NOM.SG.M	V	DAT.SG.M	ACC.SG.M
	NOM.SG.M	V	DAT.SG.F	ACC.SG.M
(8)	NOM.SG.M	V	DAT.SG.N	ACC.SG.M
	NOM.SG.M	V	DAT.PL.M	ACC.SG.M
			etc.	

In transitive utterances, there is an additional distinction based on animacy for noun phrases in the Object position of the utterance, which yields 72 types in the NOM-ACC configuration and 72 in the NOM-DAT configuration. Together, there are 360 unique utterance subtypes. As can be gleaned from the utterance types, the genitive case is not considered within the experiments because it is not part of basic German argument structures.

The choice for utterance *types* instead of *tokens* has been made because (a) the German language model is compared to a more transparent alternative for which no corpus data exist, and (b) the model is a so-called *precision model* that performs full analysis and production of a sentence. Precision models are currently too labor-intensive to build for achieving robust broad coverage, hence they cannot always be used on a large corpus. However, they are well suited for investigating the complex interplay between different linguistic constraints, so they are very good at identifying the parts of the grammar that contribute to the disambiguation task and which do not.

The use of *types* means that the model can only identify *potential* problems in language processing, whereas *tokens* would offer a more complete picture of *actual* processing complexity. It is therefore important to complement the kind of experiments as reported in this study with models that take token frequency into account, such as the information-theoretic approaches in recent psycholinguistics studies (Hale, 2003; Levy, 2008; Piantadosi et al., 2012). However, because these models have to achieve robust broad-coverage performance, they incorporate *shallow language models* that are unable to represent all of the conditional probabilities involved in language processing, and very little work exists on language production. Future research therefore needs to find a way to combine the robustness of shallow language models with precision grammars.

The Grammar and its Variant. The experiments use a grammar sample for German implemented in Fluid Construction Grammar, capable of both parsing and producing the sentences of the corpus. Besides lexical constructions, the grammar involves a Determiner-Noun Phrase construction, and two argument structure constructions (transitive and ditransitive). Apart from a Verb-Second constraint and Subject-Verb Agreement, these argument structure constructions do not include word order information, which means that the grammar does not exploit word order for disambiguating an utterance's argument structure. Lexical verbs specify whether they take Accusative or Dative Objects, and may feature selection restrictions. The grammar uses definite articles as the sole carriers of markings of the case declension system.

In the comparative experiment, the paradigm of definite articles is replaced by a different system that features more transparency in order to test our hypothesis. The variant may be an artificial one, but this paper uses a reconstruction of the Old High German system (500–1100 AD; Wright, 1906), as illustrated

Table 11.8 *The Old High German definite article paradigm.*

Case	SG-M	SG-F	SG-N	PL-M	PL-F	PL-N
NOM	<i>dér</i>	<i>diu</i>	<i>daž</i>	<i>die</i>	<i>deo</i>	<i>diu</i>
ACC	<i>dén</i>	<i>die</i>	<i>daž</i>	<i>die</i>	<i>deo</i>	<i>diu</i>
DAT	<i>demu</i>	<i>déra</i>	<i>demu</i>	<i>dém</i>	<i>dém</i>	<i>dém</i>
GEN	<i>dës</i>	<i>dëra</i>	<i>dës</i>	<i>dëro</i>	<i>dëro</i>	<i>dëro</i>

through the OHG definite articles in Table 11.8. The table does not show the instrumental case, which has disappeared from the language. Choosing the OHG system is more interesting than taking an artificially constructed variant because the comparison between the current system and its actual predecessor may help us learn something about why the OHG system evolved into its current form. One remarkable observation is that the OHG system featured twice as many forms as the current one, which begs the question why this more transparent system evolved into a more syncretic one.

11.3.2 Assessment Criteria and Results

The following subsections introduce the linguistic assessment criteria used in the experiments. The results are always extrapolated on a scale between zero and one, for which holds that the higher the score, the better the performance. For each measure, I first introduce an implementation-independent description, followed by its definition as operationalized in the experiments. The criteria themselves are thus assumed to be cognitively plausible, but the research does not commit itself to a specific implementation. In the remainder of this chapter, I use the term *language system* to refer to a specific grammatical aspect of a language, such as its system of definite articles, tense-aspect system, and so on. I will use the term *language* to refer to the whole linguistic inventory (i.e., containing the lexicon and all language systems).

11.3.2.1 Cue Reliability and Disambiguation Power Each language system within a particular language can be considered as a *cue* to help the disambiguation task of the listener. Consider for instance the sentences in example (9) and look at whether the Number value of the Subject of each sentence can be properly disambiguated.

- (9)
- a. An antelope ran away.
 - b. The man crossed the street.
 - c. The fish were biting well that day.
 - d. The antelope ran away when we tried to approach them.
 - e. The antelope ran away.

In sentence (9a), the indefinite article is the cue that the speaker can rely on for disambiguating the Subject's number, whereas the noun *antelope* (which can be both singular or plural) and the verb form *ran* did not offer a reliable cue. In the next three sentences, the listener each time relies on a different cue: the singular noun form *man* (9b), the plural verb form *were* (9c), and the plural anaphoric pronoun *them* (9a). In the last sentence, the grammar does not offer the listener any help in figuring out the Number of the Subject.

Measures. *Cue reliability* is a measure that evaluates how reliable each cue (or language system) is for disambiguation without looking at other cues. The cues of example (9) all have a low reliability, because they were only helpful in one out of five utterances. The *Disambiguation Power* of a language, however, measures how often the grammar proves itself useful when all of these cues are pitched together. In example (9), the grammar was successful in four out of five cases, despite the low reliability of each individual cue. In the experiments, these measures are applied for measuring how reliable the case declension system is for disambiguating the argument structure of an utterance, and how much disambiguation power the German grammar sample has in total.

The measures are operationalized as shown in the equations in (10) and (11). Let \mathcal{L}_i be a particular language system, and $CR(\mathcal{L}_i)$ the cue reliability of that language system. $CR(\mathcal{L}_i)$ is calculated by dividing the number of disambiguated utterances given the language system ($U_D|\mathcal{L}_i$) by the total number of utterances U . A language's disambiguation power DP is calculated in a similar way, namely by dividing the number of disambiguated utterances given all available systems in a language ($U_D|\mathcal{L}_i, \mathcal{L}_{i+1}, \dots, \mathcal{L}_n$) by the total number of utterances U .

$$CR(\mathcal{L}_i) = \frac{(U_D|\mathcal{L}_i)}{U} \quad (10)$$

$$DP = \frac{(U_D|\mathcal{L}_i, \mathcal{L}_{i+1}, \dots, \mathcal{L}_n)}{U} \quad (11)$$

Results. Figure 11.4 compares the number of ambiguous utterances using OHG (in black) and NHG (in white) when parsing and interpreting all 360 utterance types. The X-axis shows the number of utterances. The Y-axis shows four different analyses of the experiments: on the top, the number of disambiguated utterances is shown when the grammar only uses the case specification of the definite articles and the number-gender information of their head nouns ($U_D|\mathcal{L}_1$). These results form the basis for calculating the cue reliability of the OHG- and NHG-systems of definite articles. The other three analyses

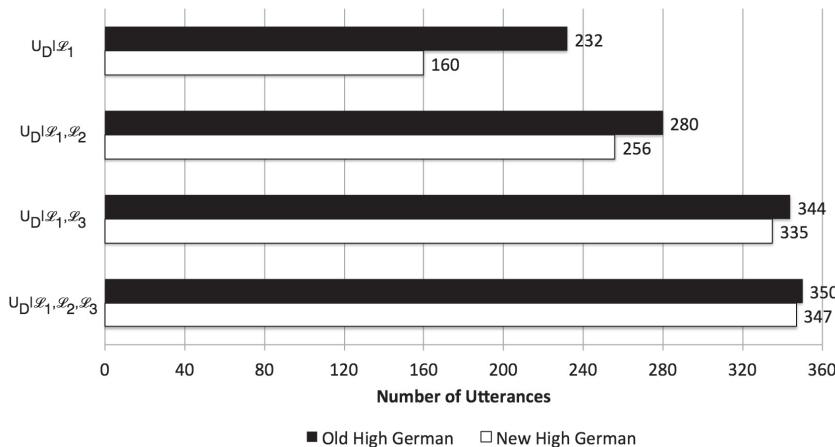


Figure 11.4 This chart compares Old High German (black) and New High German (white) in how many utterances were disambiguated during parsing in four different analyses. In all four set of results, the information carried by the determiners and the number-gender information from their head nouns (\mathcal{L}_1) were used. The first set of results show that, in isolation, the OHG determiners are more reliable for disambiguating utterances than the NHG determiners. However, as seen in the other sets of results, where the listener is also allowed to exploit other grammatical cues such as subject-verb agreement (\mathcal{L}_2) and/or selection restrictions (\mathcal{L}_3), the difference between the two systems almost disappears.

show the number of ambiguous utterances when other cues were added, such as subject-verb agreement (second set of results: $(U_D | \mathcal{L}_1, \mathcal{L}_2)$), semantic selection restrictions (third set of results: $(U_D | \mathcal{L}_1, \mathcal{L}_3)$), and both subject-verb agreement and selection restrictions (fourth set of results: $(U_D | \mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3)$). The fourth set of results forms the basis for calculating the disambiguation power of the two systems.

The results clearly show that the OHG definite articles offer a far more reliable cue for disambiguating utterances with 232 disambiguated utterance types, as opposed to 160 using the NHG articles. The most problematic utterance types for both grammars are all types that involve a plural subject and plural direct object, and all types that either involve singular-neuter arguments in nominative and accusative case, or any combination of a singular-neuter and a plural argument in nominative and accusative case. For NHG, there are additional problems for disambiguating utterance types that involve singular-feminine subjects and direct objects, or any combination of singular-feminine and plural noun phrases in nominative or accusative case.

However, when the agents are also allowed to use other cues, it becomes clear that much of the loss in cue reliability of the NHG-system can be compensated for. By also exploiting subject-verb agreement, the number of disambiguated utterance types increases from 232 to 280 for Old High German (an improvement of 13.3%) and from 160 to 256 for New High German (an improvement of 26.7%). Both grammars can profit from the fact that all types that involve a nominative singular-neuter subject with an accusative plural object (or vice versa) are now properly disambiguated. The NHG system can additionally profit from the fact that all combinations of a nominative singular-feminine subject with an accusative plural object (or vice versa) are now disambiguated as well.

The third set of results shows that semantic selection restrictions have an even bigger impact on the overall disambiguation power of the language. Here we see that OHG disambiguates 344 utterances (an improvement of 31.1% with respect to the individual cue reliability of the OHG articles) and that NHG disambiguates 335 utterances (an improvement of 48.7% with respect to the individual cue reliability of NHG articles). When exploiting both subject-verb agreement and selection restrictions in addition to the information provided by the articles and by the number-gender information of their head nouns, we see that the difference between both grammars is reduced to three utterance types, which are due to the fact that NHG does not distinguish between nominative and accusative singular-feminine arguments. The OHG-grammar now only fails to disambiguate 10 out of 360 utterance types (an improvement of 32.8% with respect to the first set of results), and the NHG-grammar now fails to disambiguate only 13 utterance types (an improvement of 52.0%).

We can now calculate the cue reliability of both systems and the overall disambiguation power of the two grammars using the equations given in (9–10). The cue reliability of the OHG-paradigm is 0.644 (i.e., speakers can rely on it for disambiguating 64.4% of the utterance types) versus only 0.444 for the NHG-system (i.e., speakers can rely on it for disambiguating 44.4% of the utterance types). This difference confirms our intuition that syncretism harms cue reliability. However, the total disambiguation power of OHG is 0.972 as opposed to 0.964 for NHG, which is a much smaller difference. Given the fact that the experiments did not take noun declension, intonation and word order preferences into account, the real difference may turn out to be even smaller, hence the increased syncretism in NHG definite articles does not seem to have harmed the language at all. This result contradicts our intuition that grammars consisting of more reliable subsystems are necessarily better at disambiguation than grammars consisting of only partially reliable subsystems. When pitched together, such systems can guarantee robust disambiguation power as long as they are sufficiently complementary to.

11.3.2.2 Processing Efficiency Processing efficiency measures the computational resources that language users need to allocate to the task of producing and parsing utterances. We have already established that syncretic forms do not lead to bigger search spaces if some form of *paradigmatic inferencing* can be used, which is achieved in the FCG implementation through the use of distinctive feature matrices. So the question then becomes: In which case is paradigmatic inferencing more efficient?

Measure. Paradigmatic inferencing requires the linguistic processor to access each cell in a morphological paradigm. The processing cost of paradigmatic inferencing can thus be measured by counting these primitive operations. Formally: let a feature matrix FM consist of a set of feature-value pairs $FM = \{FV_i, FV_{i+1}, \dot{E}, FV_n\}$. Let the processing cost PC of a feature matrix be the sum of the length of the matrix $|FM|$ (i.e., the amount of case distinctions) and the lengths of each feature-value pair, multiplied by two.

$$PC(FM) = 2 \times (|FM| + \sum_{i=1}^{|FM|} |FV_i|) \quad (12)$$

Given the processing cost of paradigmatic inferencing, we can extrapolate the *processing efficiency* of a morphological paradigm by comparing its actual cost to its maximal cost. The maximal cost of a paradigm equals the cost of a theoretical paradigm in which each cell has its own distinct marker. For German case, the most elaborate paradigm would contain 18 different forms (3 case distinction \times 2 number distinctions \times 3 gender distinctions). By dividing the actual cost by the maximal cost, we extrapolate the processing cost of a paradigm on a scale from 0 to 1. The processing efficiency of a feature matrix $E(FM)$, which is the opposite of cost, can now simply be calculated by subtracting the interpolated cost from 1. The equation is shown in example (13).

$$E(FM) = 1 - \frac{PC(FM)}{MPC} \quad (13)$$

Results. With the equation given in (12), we can first calculate the maximal cost of processing an imaginary German system of definite articles that has 18 distinct forms to serve as a baseline. That cost is 54:

$$\begin{aligned} MPC &= 2 \times (|FM| + |FV_{NOM}| + |FV_{ACC}| + |FV_{DAT}|) \\ &= 2 \times (3 + 8 + 8 + 8) \\ &= 54 \end{aligned} \quad (14)$$

The OHG-system, which only features a few collapsed cells, is 48:

$$\begin{aligned} PC(FM_{OHG}) &= 2 \times (|FM_{OHG}| + |FV_{NOM}| + |FV_{ACC}| + |FV_{DAT}|) \\ &= 2 \times (3 + 8 + 8 + 5) \\ &= 48 \end{aligned} \tag{15}$$

The New High German, which has a smaller case paradigm, has a lower cost of 40:

$$\begin{aligned} PC(FM_{NHG}) &= 2 \times (|FM_{NHG}| + |FV_{NOM}| + |FV_{ACC}| + |FV_{DAT}|) \\ &= 2 \times (3 + 6 + 6 + 5) \\ &= 40 \end{aligned} \tag{16}$$

Applying the equation in (13), the processing efficiency of the OHG-system is 0.111 compared to 0.260 for the NHG-system. In other words, because the plural distinctions have collapsed into a single cell, the linguistic processor can perform paradigmatic inferencing twice as fast with the NHG-system than with the OHG-system without any significant loss in disambiguation power. The results thus indicate that while both paradigms are equally good at eliminating ambiguities at processing time, the NHG-system is more efficient in doing so.

One of the anonymous reviewers notices that the feature matrix representation for the declension system is more favourable to the NHG system than to the OHG system. While it is true that each representation is biased, the only bias of a feature matrix is to favour smaller paradigms. What is more important than this bias, is how the members of the paradigm fill each cell. If they are not strategically opposed to each other, the benefit of a small paradigm would be outweighed by the increased ambiguity. Secondly, even if the bias of a feature matrix plays a role in the results, the experiments prove that syncretism does not automatically lead to an increase in processing cost.

11.3.2.3 Ease of Articulation One of the common assumptions about speech is that it needs to find a balance between pronunciation economy on the one hand and intelligibility on the other. In other words, speakers are assumed to prefer forms that require the least *articulatory effort* while at the same remaining distinct enough to be properly distinguished from other forms in the language.

Measure. A popular measure for assessing articulatory effort is based on tracking the movements of articulators (such as the lips, tongue, and uvula) when pronouncing speech sounds (Perkell et al., 2002). The experiments presented in this chapter do not involve a real speech system, but simulate phonological sounds using a method proposed by Stevens (2002). More

Table 11.9 *This Table illustrates the discrete representation of the NHG definite articles die and das in sets of distinctive features per phoneme. Irrelevant features have no value. Diphthongs are represented as two separate phonemes.*

Phonemes	die		das		
	d	i:	d	a	s
syllabic	—	+	—	+	—
continuant	—		—		+
sonorant	—		—		—
nasal	—		—		—
voice	+		+		—
anterior	+		+		+
coronal	+		+		+
lateral	—		—		—
high	—	+	—	—	—
low		—		+	
back	—	—	—	+	—
rounded		—		—	
long		+		—	

specifically, each definite article construction contains a discrete representation of the phonemes required for pronouncing the article, with each phoneme described by a set of binary distinctive features (such as [voice +] or [nasal -]). The distinctive feature sets used for representing the sounds of New High German articles are taken from Russ (1994); sets for Old High German are reconstructed based on descriptions by Wright (1906). Table 11.9 shows the distinctive feature sets for the NHG articles *die* and *das*.

Using this representation, we can calculate articulatory effort in terms of how many feature-value pairs are changed when moving from one phoneme to the next. Following a widespread tradition in computational linguistics, this chapter operationalizes these changes as an *edit distance* with a few minor adaptations. Articulatory effort A is thus calculated as follows. Let the cost of moving from one set of distinctive features S_i describing a phoneme to the next one S_{i+1} be $c_f(S_i, S_{i+1})$. A is the sum of all costs until the last sound has been reached, with k as the number of phonemes, as shown in (17).

$$A = \sum_{i=1}^{k-1} c_f(S_i, S_{i+1}) \quad (17)$$

As already said, each cost $c_f(S_i, S_{i+1})$ is measured as an edit distance, where we add the amount of deletions or insertions (i.e., the non-shared features $F_n = \{f_1, \dots, f_n\} = S_i \Delta S_{i+1}$) to two times the amount of substitutions (i.e., the

amount of shared features $F_s = \{f_1, \dots, f_m\} = S_i \cap S_{i+1}$ whose values are different in the two sets). For example, the articulatory effort for *die* is 14 (10 non-shared features + 4 for two shared features with a different value), whereas the effort for *das* is 28 (14 for moving from [d] to [a] and 14 for moving from [a] to [s]). So *die* is more economic than *das*.

Ease of articulation can then be inferred by comparing the average effort of OHG articles and NHG articles against a baseline of maximal effort. Here, I take maximal effort to be 78, which would be the effort of an imaginary article consisting of four phonemes (which corresponds to the size of the largest forms in the comparison, namely OHG *dēmu* and *dēru*) that are maximally distant from each other. Using the maximal effort, we can interpolate the articulatory effort of an article on a scale of zero to one by dividing its actual cost by the maximal effort. For instance, the interpolated effort of the NHG article *die* is 0.189 (= 14/78). Ease of articulation is the opposite measure of articulatory effort and is calculated by taking the difference of 1 and the interpolated effort. The ease of articulation of *die* is therefore 0.811.

Results. Table 11.10 shows two results for each article in the OHG and NHG paradigms: articulatory effort on the left and its ease of articulation on the right. In the OHG paradigm, the average ease of articulation is 0.668 (calculated by counting the ease of articulation of each cell in the paradigm and then dividing it by the number of cells). The NHG paradigm, on the other hand, has an ease of articulation of 0.733, so the NHG articles require on average less effort to pronounce.

Looking at the results in more detail, we see that the culprits for the lower score of OHG are the dative-singular forms *dēmu* and *dēru*, who are the most “expensive” articles as they consist of four phonemes. Interestingly, these expensive forms have undergone phonological erosion and have become more economic in NHG. Another interesting observation is that also the three “cheapest” forms in the OHG paradigm (*die*, *deo*, and *diu*), which all three end in a diphthong, have eroded into an even more economic form *die* in NHG (now ending in a long vowel).

This begs the question why the other articles (surviving as *der*, *den*, and *das* in NHG), which require more effort than *die*, were not further reduced by phonological erosion. The answer is semantic ambiguity. Recall from the earlier results that the three nominative and accusative plural forms in OHG did not contribute anything to the language’s disambiguation power. Likewise, the shorter dative forms in the NHG system, even though increasing the syncretism in the paradigm, were harmless for the language’s disambiguation power. However, if, for instance, *der* and *den* were to collapse, the NHG system would not contribute anymore to the disambiguation of nominative from accusative arguments in an utterance.

Table 11.10 *The articulatory effort for each definite article on the left, and its ease of articulation to the right. The ease of articulation has increased from OHG to NHG through phonological reduction in the dative-singular forms and by shifting the diphthongs of the nominative and accusative plural forms to a long vowel.*

Old High German											
Singular						Plural					
		Masc	Neut	Fem		Masc	Neut	Fem			
MOM	dér	24	0.6757	26	6.6486	20	0.7297	18	0.7568	20	0.7297
	dén			daz̄		die	diu		die		deo
ACC	dén	24	0.6757	26	0.6486	18	0.7568	18	0.7568	20	0.7297
	dēmu			dēmu		dēru	den		den		den
DAT	dēmu	40	0.4595	40	0.4595	40	0.4595	24	0.6757	24	0.6757
New High German											
Singular						Plural					
		Masc	Neut	Fem		Masc	Neut	Fem			
NOM	der	24	0.6757	26	0.6486	14	0.8108	14	0.8108	14	0.8108
	den			das		die	die		die		die
ACC	den	24	0.6757	26	0.6486	14	0.8108	14	0.8108	14	0.8108
	dem			dem		der	den		den		den
DAT	dem	24	0.6757	24	0.6757	24	0.6757	24	0.6757	24	0.6757

11.3.2.4 Auditory Distinctiveness The payoff for “articulatory laziness” is that the listener needs more auditory precision in order to understand the speaker, which is measured in terms of *auditory distinctiveness*. That is, the more distinct a form is from other forms, the easier it is to recognize for the listener.

Measure. We can use the same distinctive feature representation for calculating the distance between an observed form and its nearest phonological neighbors. As the identification of parts-of-speech is considered to be a solved problem in computational linguistics, I assume here that the agents are able to recognize an article when they perceive one, so only other articles are taken into consideration for measuring auditory distinctiveness. First, the phonemes of two articles are mapped onto each other using their discrete representation as just described earlier, which is here illustrated for the NHG-articles *die* and *das*:

Table 11.11 This Table shows for each article its articulatory distinctiveness with respect to its nearest phonological neighbor on the left, and the same measure interpolated on a scale from zero to ten to the right.

Old High German												
Singular						Plural						
		Masc	Neut	Fem		Masc	Neut	Fem				
MOM	2	<i>dér</i>	<i>daz</i>	<i>diu</i>		<i>die</i>	<i>diu</i>	<i>deo</i>				
		0.0192	6	6.0577	6	0.0577	6	0.0577	6	0.0577	8 0.0769	
ACC	2	<i>dén</i>	<i>daz</i>	<i>die</i>		<i>die</i>	<i>diu</i>	<i>deo</i>				
		0.0192	6	6.0577	6	0.0577	6	0.0577	6	0.0577	8 0.0769	
DAT	2	<i>dëmu</i>	<i>dëmu</i>	<i>dëru</i>		<i>den</i>	<i>den</i>	<i>den</i>				
		0.0192	2	0.0192	2	0.0192	2	0.1923	2	0.0192	2 0.0192	
New High German												
Singular						Plural						
		Masc	Neut	Fem		Masc	Neut	Fem				
NOM	2	<i>der</i>	<i>das</i>	<i>die</i>		<i>die</i>	<i>die</i>	<i>die</i>				
		0.0192	8	0.0769	14	0.1346	14	0.1346	14	0.1346	14 0.1346	
ACC	2	<i>den</i>	<i>das</i>	<i>die</i>		<i>die</i>	<i>die</i>	<i>die</i>				
		0.0192	8	0.0769	14	0.1346	14	0.1346	14	0.1346	14 0.1346	
DAT	2	<i>dem</i>	<i>dem</i>	<i>der</i>		<i>den</i>	<i>den</i>	<i>den</i>				
		0.0192	2	0.0192	2	0.0192	2	0.1923	2	0.0192	2 0.0192	

$$(18) \quad \begin{array}{c|c|c} d & i: & - \\ \hline d & a & s \end{array}$$

The total distance D between two forms is calculated as the sum of all the distances between two mapped sets of distinctive features $d_f(S_i, S_{i'})$:

$$D = \sum_{i=1}^k d_f(S_i, S_{i'}) \quad (19)$$

The distance function d_f is calculated in the same way as the cost function c_f described in the previous section. In our example, the distance between *die* and *das* is 18 (0 for the shared phoneme [d], 8 for the distance between [i:] and [a], and 10 for all the nonshared features between the zero pronunciation of *die* and [s] of *das*). The articulatory distinctiveness can be interpolated to a scale between zero and one by dividing it by the maximal distance. Here, I take the maximal distance between two articles to be 104, which would be the distance

between two imaginary morphemes of four phonemes each with a maximal distance between each phoneme.

Results. The results for each article are shown in Table 11.11, with the actual distance between an article and its nearest neighbor on the left and the interpolated auditory distinctiveness on the right. The first thing to notice is that all articles have at least one very close neighbor. Even the most distinct article (*die* in the NHG-system) still only scores 0.135 on a scale of zero to one. The results thus suggest that ease of articulation is more important than auditory distinctiveness, and that our auditory perception is well capable of distinguishing only minimally diverging sound patterns. On average (by taking the sum of each cell and dividing it by the total number of cells), the NHG system scores slightly better than the OHG system with an auditory distinctiveness of 0.077 versus 0.043.

Despite the small differences, we can still see some clusters appearing if we look at the results in more detail. Figure 11.5 presents four spider charts that each take one article as their center and then show the distances between that article and the other forms in its paradigm. The figure shows spider charts for the OHG articles *die* and *dēru* on the left, and two charts for their corresponding NHG forms *die* and *der* on the right.

On the top left we see the distances between OHG *die* (the center value 0) and the other OHG articles. As can be seen, the forms *diu* and *deo* are closer to the center than the other articles are, which means they are harder to distinguish from each other. In the NHG system, all three forms have collapsed into the article *die*, which can be distinguished from the other NHG articles more easily, as shown in the spider chart on the top right. Given that *die*, *deo*, and *diu* in OHG anyway did not contribute to the language's disambiguation power, the loss of an unnecessary distinction has also made the task of perceiving the articles easier.

A second cluster of forms that are hard to distinguish from each other in the OHG paradigm is shown in the spider chart on the bottom left of Figure 11.5, which illustrates the distances between the article *dēru* in the center and the other forms in the paradigm. The chart shows that *dēmu* is the closest article, followed by *dēr*. Both *dēru* and *dēmu* have undergone phonological reduction, but the result is a new cluster in NHG (shown on the bottom right) in which *der*, *den*, and *dem* are very close to each other. From the viewpoint of auditory distinctiveness, this is a less than optimal arrangement and it may seem surprising that such minimal differences have been able to survive for centuries in the German language. Here again it seems that semantic ambiguity plays the referee in deciding whether or not to uphold a case distinction: collapsing *der* and *dem* would result in ambiguities in any NOM-DAT pattern involving singular-masculine arguments, and as already said, collapsing *der* and *den* would make the system useless for distinguishing nominative from accusative arguments.

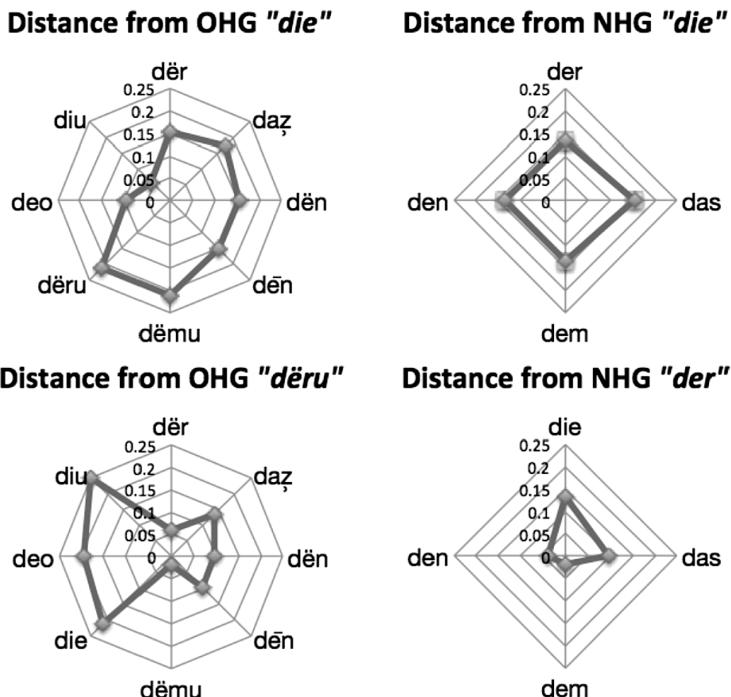


Figure 11.5 These spider charts each take an article at their center and then show the interpolated distances between that article and other forms of the same paradigm. The center articles are OHG *die* (top left) and *dēru* (bottom left) and their corresponding forms in NHG *die* (top right) and *der* (bottom right). When comparing the spider charts for OHG and NHG *die*, we see that a cluster of close OHG forms (*die*, *diu*, and *deo*) has collapsed into a single form in NHG, which has improved the auditory distinctiveness of the paradigm. When comparing OHG *dēru* to NHG *der*, we see that the OHG forms of *dēmu* and *dēru* have undergone phonological reduction, which results in three close neighbors in the NHG system: *der*, *den*, and *dem*. Despite their low auditory distinctiveness, these forms have been maintained in order to uphold the language's disambiguation power.

11.4 Discussion and Conclusions

In recent years, there has been a growing interest in the puzzle of ambiguity in natural languages. The mainstream view, most outspokenly defended by Chomskyan linguistics, ambiguity was assumed to be detrimental for communication and even as evidence that language is therefore not well designed for serving its communicative purpose. Cognitive-functional linguists, however, claim that ambiguity serves an important communicative function. These claims are

increasingly being backed up by rigorous formalization and experiments. For example, Piantadosi et al. (2012) provide information-theoretic arguments and formalization to support the claim that ambiguity makes inferential communication systems more efficient.

This chapter supports this new movement through a case study on the German declension system. After reviewing the empirical observation of the system, which shows that German noun phrases typically choose either the article or the adjective for carrying case-number-gender inflection, it has proposed a formalization of the German declension system as a distinctive feature matrix in Fluid Construction Grammar. Such a feature matrix allows paradigmatic inferencing during processing, which allows case syncretisms to be efficiently processed as long as the case forms are still in functional opposition of each other. The chapter then suggested that, in terms of a number of communicative assessment criteria, the current German declension system is indeed more efficient than its Old High German predecessor, which had more transparent form-meaning mappings. More specifically, ambiguity of case forms has allowed the German declension system to reduce its size in half without harming the Disambiguation Power of the language.

One important warning is that it would be wrong to conclude that the New High German system is more efficient than the Old High German one: the novel methodology proposed in this chapter has plugged the Old High German system of definite articles into an otherwise New High German grammar model. In other words, the comparison only shows what the performance of the Old High German declension system would have been if it would still be part of German today. When reviewing the linguistic facts, it turns out that the Old High German definite articles only grammaticalized into articles later on. Before, they functioned as demonstratives that appeared without a head noun, which meant that they did not have access to the Gender and Number information of the head noun. In such a “linguistic ecosystem” the individual cue reliability of these case forms is of greater importance than in the current one, where there is a more established noun phrase. In other words, it is more correct to say that the NHG declension system is more adapted to (or efficient in) the constraints of the current German grammar, whereas the OHG system seems to be more adapted to the constraints of the German grammar as it was in the OHG time period.

The result is that we get an interesting trade-off between transparency between meaning and form on the one hand and processing efficiency on the other. Whenever the language can rely on other linguistic cues for helping the listener’s disambiguation efforts, an increase in local ambiguity can be expected. Future research will have to provide more evidence that this is indeed the case.

Acknowledgments

The research reported in this essay has been conducted at and funded by the Sony Computer Science Laboratory Paris. The author wishes to thank his colleagues for their support and constructive feedback, particularly Luc Steels and Katrien Beuls. The computational implementation for German would not have been possible without the help of my German informants Sebastian Höfer, Vanessa Micelli and Michael Spranger, and without my former colleagues Joris Bleys, Joachim De Beule, Martin Loetzsch, and Pieter Wellens, who helped make FCG a powerful tool for exploring issues in constructional language processing. I also thank Stefan Müller for his insights on formal approaches to German, and the many researchers I interacted with when presenting this work. Special thanks also go to Aline Villavicencio and Thierry Poibeau for their patience and for offering me the chance to publish this work in a superb collection of papers, and to the reviewers for making insightful suggestions and comments. All remaining errors are of course my own.

References

- M. Bierwisch. Syntactic features in morphology: General problems of so-called pronominal inflection in German. In *To Honour Roman Jakobson*, pages 239–270. Mouton De Gruyter, Berlin, 1967.
- James Blevins. Syncretism and paradigmatic opposition. *Linguistics and Philosophy*, 18:113–152, 1995.
- David Carter. Efficient disjunctive unification for bottom-up parsing. In *Proceedings of the 13th Conference on Computational Linguistics*, pages 70–75. ACL, 1990.
- Noam Chomsky. On phases. In Robert Freiden, Carlos P. Otero, and María Luisa Zubizarreta, editors, *Foundational Issues in Linguistic Theory: Essays in Honor of Jean-Roger Vergnaud*, pages 133–166. MIT Press, Cambridge MA, 2008.
- H. Clahsen, M. Hadler, S. Eisenbeiss, and I. Sonnenstuhl-Henning. Morphological paradigms in language processing and language disorders. *Transactions of the Philological Society*, 99(2):247–277, 2001.
- Ann Copestake. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, 2002.
- B. Crysmann. Syncretism in german: A unified approach to underspecification, indeterminacy, and likeness of case. In S. Müller, editor, *Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar*, pages 91–107, Stanford, 2005. CSLI.
- Walter Daelemans and Antal Van den Bosch. *Memory-Based Language Processing. Studies in Natural Language Processing*. Cambridge University Press, Cambridge, 2005.
- Mary Dalrymple, Tracy Holloway King, and Louisa Sadler. Indeterminacy by underspecification. *Journal of Linguistics*, 45:31–68, 2009.
- M. Daniels. On a type-based analysis of feature neutrality and the coordination of unlikes. In *Proceedings of the 8th International Conference on HPSG*, pages 137–147, Stanford, 2001. CSLI.

- Sonja Eisenbeiss, Suzanne Bartke, and Harald Clahsen. Structural and lexical case in child German: Evidence from language language-impaired and typically-developing children. *Language Acquisition*, 13:3–32, 2005/2006.
- D.P. Flickinger. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1):15–28, 2000.
- J.T. Hale. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2):101–123, 2003.
- W. Heinz and J. Matiasek. Argument structure and case assignment in German. In J. Nerbonne, K. Netter, and C. Pollard, editors, *German in Head-Driven Phrase Structure Grammar*, pages 199–236. CSLI, Stanford, 1994.
- R.J.P. Ingria. The limits of unification. In *Proceedings of the 28th Annual Meeting of the ACL*, pages 194–204, 1990.
- L. Karttunen. Features and values. In *Proceedings of the 10th International Conference on Computational Linguistics*, Stanford, 1984.
- Sterre Leufkens. *Transparency in Language: A Typological Study*. LOT, Utrecht, 2015.
- Robert Levine, Thomas Hukari, and Michael Calcagno. Parasitic gaps in english: Some overlooked cases and their theoretical consequences. In Peter W. Culicover and Paul M. Postal, editors, *Parasitic Gaps*. MIT Press, Cambridge MA, 2001.
- Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.
- Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA, 1999.
- Gereon Müller. Remarks on nominal inflection in German. In Ingrid Kaufmann and Barbara Stiebels, editors, *More than Words: A Festschrift for Dieter Wunderlich*, pages 113–145. Akademie Verlag, Berlin, 2002.
- Stefan Müller. An HPSG-analysis for free relative clauses in german. *Grammars*, 2(1): 53–105, 1999.
- Stefan Müller. Case in German – towards and HPSG analysis. In Tibor Kiss and Detmar Meurers, editors, *Constraint-Based Approaches to Germanic Syntax*. CSLI, Stanford, 2001.
- J. Perkell, M. Zandipour, M. Matthies, and H. Lane. Economy of effort in different speaking conditions. i. a preliminary study of intersubject differences and modeling issues. *J. Acoust. Soc. Am.*, 112:1627–1641, 2002.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122:280–291, 2012.
- Geoffrey Pullum and Arnold Zwicky. Phonological resolution of syntactic feature conflict. *Language*, 62(4):751–773, 1986.
- Allan Ramsay. Disjunction without tears. *Computational Linguistics*, 16(3):171–174, 1990.
- C. Russ. *The German Language Today. A Linguistic Introduction*. Routledge, London, 1994.
- I.A. Sag. Coordination and underspecification. In J. Kom and S. Wechsler, editors, *Proceedings of the Ninth International Conference on HPSG*, Stanford, 2003. CSLI.
- Luc Steels. Constructivist development of grounded construction grammars. In Walter Daelemans, editor, *Proceedings 42nd Annual Meeting of the Association for Computational Linguistics*, pages 9–19, Barcelona, 2004.

- K.N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.*, 111:1872–1891, 2002.
- Remi van Trijp. A design pattern for argument structure constructions. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*. John Benjamins, Amsterdam, 2011a.
- Remi van Trijp. A design pattern for argument structure constructions. In Luc Steels, editor, *Design Patterns in Fluid Construction Grammar*, pages 205–235. John Benjamins, Amsterdam, 2011b.
- Remi van Trijp. Linguistic selection criteria for explaining language change: A case study on syncretism in German definite articles. *Language Dynamics and Change*, 3:105–132, 2013.
- B. Wiese. Iconicity and syncretism. on pronominal inflection in Modern German. In R. Sckmann, editor, *Theoretical Linguistics and Grammatical Description*, pages 323–344. John Benjamins, Amsterdam, 1996.
- Susanne Winkler, editor. *Ambiguity: Language and Communication*. Walter De Gruyter, Berlin, 2015.
- Joseph Wright. *An Old High German Primer*. Clarendon Press, Oxford, 2nd edition, 1906.
- Dieter Wunderlich. Der unterspezifizierte Artikel. In Karl Heinz Ramers Dürscheid and Monika Schwarz, editors, *Sprache im Fokus*, pages 47–55. Niemeyer, Tübingen, 1997.

Index

- Accusative, 109, 292, 293, 295, 297–298, 302, 305–306, 310–311, 313
Activation, 9, 34, 45, 54, 63–68, 276
Agreement, 12, 36, 43, 60, 106, 108, 290, 293, 295–297, 302, 305–306
Altruism, 227, 234–236, 239–240, 242, 259–260
Altruistic, 17, 228, 233, 235–237, 239, 240–242
Alzheimer, 12, 15, 82, 84, 95
Anterior temporal lobe, 66, 68–69
Aphasia, 55–56, 261
Argument Structure, 12, 134, 295, 302, 304
Articulation, 59, 188, 257, 262, 308, 310–311, 313
Articulatory, 61, 113, 277, 308–312
Attributes, 40, 81–82, 86–92, 94–95, 161, 178
Auditory, 32, 59, 61, 71, 300, 311, 313–314

Baldwin effect, 258, 279–281
Bayes, 12, 16–17, 90, 160, 170, 185–187, 190–191, 199–200, 206, 213, 219
Bayes' theorem, 268
Bayesian, 12, 16–17, 170, 185–187, 190–191, 199–200, 206, 213, 219, 256, 266–272, 273–275, 282
Bayesian iterated learning, 267
Bayesian segmentation, 16, 185–187, 190, 199, 206, 219
Bigram, 192–195, 199–201, 206, 221
Bipolar, 15, 81, 83–84, 86, 88–91, 93
Broca's area, 55–56, 61–62

Case declension, 300, 302, 304
Child-directed speech, 187, 196, 203, 208–209, 211–212, 217
Child-produced speech, 11
CIPP, 178–180
Clause-initial prepositional phrase, 178–180
Coevolution, 249, 262, 279–280, 283
Competition, 17, 114, 167, 228, 235, 262
Compositionality, 17, 122, 241, 266, 271, 282

Constrained inference, 194–195, 203, 205–206, 217
Cooperation, 17, 235, 239–242, 247
Coreference, 136–142, 145–148
Cortex, 56, 61–62, 66, 68–72, 94
Cross-linguistic, 185–187, 196, 200, 289
Cue reliability, 300, 303–304, 306, 315
Cultural evolution, 232, 256–258, 260–262, 264, 266–267, 269, 271–276, 278–279, 281–283

Dative, 109–110, 290–292, 295–299, 302, 310–311
Decayed Markov Chain Monte Carlo, 195
Deception, 230, 232, 234, 248
Declension, 18, 289–293, 298, 300–302, 304, 306, 308, 315
Decoupling, 227, 229–232, 234, 237–238, 243
Deep learning, 73
Deficit, 15, 82, 84, 91–92, 94–96
Definite articles, 290–292, 295, 302–307, 309, 315
Dementia, 15, 81–82, 84, 94–96
Dependency, 39, 41, 44–45, 132–133, 138, 140–141, 170–171
Diagnosis, 12, 15, 81–82, 84–86, 96
Dirichlet Process, 191, 193
Disambiguation power, 290, 300, 303–306, 308, 310, 313–314
Disjunctive, 293, 295, 296, 298
Disorder, 81, 83–85, 88–93
Distributional modeling, 5, 7, 16, 121, 124, 126, 130, 140, 142, 145
Distributional Semantic Model, 13, 131
Distributional Semantics, 131
Downstream evaluation, 185, 207–209, 211, 213, 217–218
Dream, 15, 81–82, 85–86, 88, 90–94, 181

Education First, 165, 182
EF-Cambridge Open Language Database, *see* EFCAMDAT

- EFCAMDAT, 16, 159–160, 164–168, 173, 181–182
 Elderly, 15, 82, 95
 Electroencephalography, 56, 94
 Emergence, 14, 17, 228, 251, 265–266, 282
 Entailment, 118–121, 136–139, 142, 148
 ERP (event-related potential), 31
 Event Schemas, 120, 135
 Event-related potential, 31
 Evolutionary, 231–233, 239, 242, 256–258, 267, 269, 278, 280
 Exaptation, 233–234, 267, 280, 283
 Extrinsic evaluation, 209, 218–219
 Eye movement, 35, 37
 F-score, 198–202, 206, 210–212, 216–218
 Feature Norms, 66, 122, 134
 Feature selection, 178, 302
 Fluid Construction Grammar, 297, 302, 315
 French, 8, 103, 104–108, 111, 117, 163–164, 166, 188
 Functional magnetic resonance imaging, 56
 Gene, 13, 230, 256–288
 Generalized Event Knowledge, 121
 Generative modeling, 69, 70, 191, 195, 201, 213, 214
 Genetic, 232, 256–288
 Good-Enough Theory, 34
 Grammar induction, 135
 Grammatical relation, 33, 165, 189
 Grammaticalisation, 315
 Graph, 12, 15, 42, 47–48, 81–87, 93–98, 136
 Gricean, 230, 241, 244, 248
 Hearer, 15, 101, 227, 230, 241–250
 Hierarchical Dirichlet Process, 193
 Hunter-gatherer, 235–236, 250
 Idealized inference, 194, 195, 199, 201, 202, 209, 215
 Idiom, 7, 34, 38, 45, 46, 48, 104, 111
 Imitation games, 275
 Implicature, 227, 241, 243–250
 Implicit, 16, 17, 101, 103–106, 134, 146, 148, 227, 231, 241, 243–251, 271, 273
 Incremental processing, 15, 27–31, 35, 43, 46, 49, 50, 72
 Infant speech perception, 187, 188, 195, 215
 Inference, 3, 48–50, 66, 108, 118–119, 145, 187–191, 194–196, 198–199, 201–206, 210, 215, 217, 245, 246
 Inferior frontal gyrus, 55, 56, 60, 61, 62
 Information Gain, 162, 177
 Interpretation, 15, 27–38, 46–50, 60, 65, 88, 192, 240, 241, 245–246, 274, 279
 Intrinsic evaluation, 207, 208, 218, 219
 Iterated learning, 256, 257, 262, 266, 267, 282
 Joint modeling, 207, 208, 209, 212, 213, 215, 216, 219
 Kohonen network, 276
 L1, 16, 86–87, 159, 161, 164, 167, 172–173, 176–177, 181
 L2 (*see* second language), 16, 86–87, 159–160, 159, 162, 164–165, 172–176, 180–181
 Language acquisition, 5, 10–12, 102, 108, 181, 185, 189, 219, 260, 261, 273, 276, 279
 Language evolution, 13, 227, 228, 229, 231–240, 256, 279
 Language identification, 15, 159–184
 Language modeling, 8, 69, 71, 72, 105, 111, 162, 191, 193, 302
 Language violations, 34, 35, 36, 37, 60
 Learner corpus, 159, 164
 Learner proficiency, 16, 159, 163–167, 172–182
 Lexical preference, 162, 178
 Lexical semantics, 28
 Lexicon, 12, 28, 29, 49, 54, 60, 63, 85, 176, 187, 190, 193, 199, 211–218, 303
 Lingueme, 260, 261
 Linguistic selectionism, 306
 Lobe, 61, 62, 64, 66, 67, 68, 69, 72, 132
 Longitudinal data, 11, 16, 159, 165, 181, 182
 Machiavellian, 228–229, 236–237
 Machine learning, 4–8, 53, 72, 90, 112, 117, 136, 143, 159, 160, 170–173, 181, 185
 Machine translation, 4, 7, 101, 103, 105, 107, 111, 114, 117, 119
 Machines, 6, 115
 Manipulation, 123, 127, 227, 239–240, 248–249, 257
 Markov Chain, 195, 268
 Matrices, 297, 298, 300, 307
 MEG (magnetoencephalography), 56–58, 60, 66, 67, 71
 Memory, 9, 10, 15, 31, 34, 35, 47–49, 61–62, 67, 69, 82, 85, 91–92, 94, 120
 Mental Lexicon, 60, 63
 Minimal semantics, 230, 231, 243
 Mitteilungsbedürfnis, 239, 242, 243, 247
 Model evaluation, 43, 48, 127, 129, 141, 143, 146, 164, 165, 187, 213, 217, 249

- Modelling, 275, 279
Monosyllabic, 197, 202, 203, 206, 209, 212
Morphology, 11, 31, 39, 109, 116, 175, 181, 186, 196, 204, 205–207
Mutualism, 235, 239–240
Mutualistic, 236, 240, 242, 247, 248

Narrative event chains, 135, 136, 137, 141
Narratives, 55, 68, 69, 136, 137, 142, 163, 164, 167
Native Language Identification, 15–16, 159, 163
Natural Language Processing, 3, 15, 27–28, 84, 101, 103, 108–109, 111–112, 114, 119, 150, 159, 289
Naturalistic stimulus, 11, 53, 54, 70, 72, 96, 217
Neo-Davidsonian, 119, 124, 131, 137, 139–140, 142
Neo-davidsonian event semantics, 119, 124, 131, 137, 139, 140, 142
Network, 12, 15, 49, 55, 61, 68, 71, 72, 120, 266, 267, 271, 296
Neural, 11, 31, 36, 61, 62, 64–66, 70–73, 268, 269, 273, 277
Neural model, 13, 71, 72, 266, 267, 271
Neuroimaging, 5, 53, 60, 62, 72, 188
Neurolinguistics, 28, 31, 32, 34, 50, 57
Ngram, 161, 297
Niche construction, 256, 279
Nominal construction, 39–42
Nomative, 109, 290–293, 295, 298–299, 305, 306, 310, 313
Normalized-segmentation entropy, 201

Online processing, 189, 195
Orthography, 126, 186, 197, 203, 211, 212, 216, 218, 219
Oversegmentation, 190, 203, 204, 205, 206

Parsimony bias, 190
Parsing, 4, 9, 27–52, 294, 295, 300, 302, 304, 307
Part of speech, 112, 169, 174
Pathological, 12, 13, 84, 90, 91, 92, 94
Phenotypic plasticity, 278, 279
Phoneme, 35, 197, 215, 260, 269, 275–277, 309, 310–313
Phonological, 14, 29, 32, 257, 277, 308, 310–314
Population, 14, 93, 95, 160, 255, 257, 258–262, 269, 271, 275, 277, 280, 281
POS (*see* part of speech), 11, 112, 169
Poverty of the stimulus, 101, 102, 108

Pragmatic, 10, 27, 103, 104, 105, 241, 245, 246, 247
Presupposition, 118, 148, 227, 243, 244, 247, 248, 249, 250
Primate, 234–243
Probability, 117
Processing effort, 295
Production rule, 161, 162, 165, 170, 172, 175–178
Property, 39–46
Property Grammars, 39
Prosocial, 17, 227–228, 234–235, 239, 250
Prosody, 27, 28, 30–32, 96
Proto-agent, 119–121, 123–124
Proto-patient, 119–121, 123
Proto-roles, 122, 140, 148
Psycholinguistics, 14, 15, 27–28, 30, 34, 54, 118, 120, 122, 133, 295, 300, 302
Psychosis, 81–84, 88, 90–91, 96
Psychotic, 81–85, 88, 90, 92–94

Reading, 31, 34–35, 53–55, 57–59, 61, 63, 65, 70–71
Reasonable errors, 185, 204–206, 211
Recency bias, 190, 195, 196
Reciprocal, 44, 235–236, 240, 262
Reciprocal altruism, 235, 236, 239
Regression, 65, 67, 162
Replicator, 260, 262, 266

Schizophrenia, 81–85, 88, 90–94, 96
Second language, 103, 159–160, 181
Second language acquisition, 103, 159, 160, 181
Second language learners, 103, 110, 111
Second language writing, 160
Selection, 17, 227, 232, 238, 239, 258–261, 267–269, 272, 275, 278–279, 282
Selectional restriction, 133, 138, 140, 141
Self-organisation, 256, 259, 275–278, 282
Semantic composition, 34, 61, 66–68, 122, 241
Semantic Memory, 62, 82, 94, 120, 122, 123, 142
Semantic role labeling, 119, 120, 130
Semantic roles, 118–119, 121, 123–125, 132, 133, 136–138, 145, 148
Semantic unification, 62, 67
Semanticity, 227, 229–232, 234, 237–238, 243
Sense, 7–8, 121
Signal, 37, 54, 57, 68, 186, 198, 229–230, 232–235, 243, 273–274
Single word reading, 54, 61, 64, 68
Slide whistle experiment, 273

- Species, 229, 231–238, 257, 267
 Speech segmentation, 185–224
 STG, 55–56, 61–62
 Stimuli, 31–32, 37, 55, 57, 60, 64, 66–67, 72, 83, 122, 124, 133, 272–273
 Story reading, 70
 Stress cues, 208–213, 218
 Support vector machine, 160, 163, 165, 170–172, 163
 Syllable, 108, 188–192, 196–197, 201–202, 206, 208–217, 274
 Symptom, 56, 82, 88–96
 Syncretism, 18, 289–318
 Syntactic structure, 15, 27–33, 49, 69, 148, 178, 188
 Syntax, 11, 27, 27–52, 62, 68–73, 109, 112, 116–117, 132–138, 148, 263–266
 Syntax-based distributional semantic models, 132–138, 148
 Thematic fit, 133–134, 137
 Thematic Roles, 16, 34, 118–121, 123, 120, 121, 123, 128–131, 133, 137–138, 148
 TOEFL, 164
 Token, 30, 197, 199–202, 206, 209–211, 217–218, 302
 Training, 7, 65–66, 84, 105, 108–112, 143, 168, 198
 Translation, 4–7, 101–117
 Translator, 101, 103–107, 113, 117
 Transmission, 229, 260–270, 272, 274–275
 Treebank, 7, 169
 Undersegmentation, 190, 203–204, 206
 Understanding, 1, 6, 9, 35, 53, 55–56, 68, 72, 96, 103, 113, 115
 Unification, 49, 61–62, 67, 297–300
 Unigram, 191–193, 199, 201, 211
 Uniqueness, 13, 39, 41–42, 234
 Unlexicalized dependencies, 170, 175
 Unsupervised learning, 131, 135, 148, 185–186, 198
 Utility-based metrics, 145, 185
 Vector Space Model, 64, 66, 69, 171
 Verb entailment, 119–121, 136–139, 142, 148
 Verb-specific inferences, 118
 Verbal semantics, 32–46, 64–67, 118–158, 265, 290, 302
 Vocabulary, 10, 63, 110, 167
 Vowel, 108, 189, 197, 261, 275–279, 282
 Wernicke's area, 55, 56, 61
 Wikipedia, 66, 164
 Word embedding, 7, 54, 64, 66, 69
 Word learning, 213, 218
 Word meaning, 7, 8, 14, 34, 54, 64, 187, 188, 213
 Word sense disambiguation, 7, 299–315