

Oxford
LINGUISTICS

edited by James P. Blevins
and Juliette Blevins

Analogy in Grammar

Form and Acquisition

competence patterns
type frequency cross-entropy
words vs. rules cognition
gestalt
paradigms feedback
exemplars word structure

Analogy in Grammar

This page intentionally left blank

Analogy in Grammar: Form and Acquisition

EDITED BY

JAMES P. BLEVINS AND JULIETTE BLEVINS

OXFORD
UNIVERSITY PRESS

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.

It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi

Kuala Lumpur Madrid Melbourne Mexico City Nairobi

New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece

Guatemala Hungary Italy Japan Poland Portugal Singapore

South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press

in the UK and in certain other countries

Published in the United States

by Oxford University Press Inc., New York

© 2009 organization and editorial matter James P. Blevins and Juliette Blevins

The moral rights of the author have been asserted

Database right Oxford University Press (maker)

© 2009 the chapters their various authors

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, or under terms agreed with the appropriate reprographics rights organization. Enquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above

You must not circulate this book in any other binding or cover and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Data available

Typeset by SPI Publisher Services, Pondicherry, India

Printed in Great Britain

on acid-free paper by

the MPG Books Group

ISBN 978-0-19-954754-8

1 3 5 7 9 10 8 6 4 2

Contents

<i>Notes on Contributors</i>	vii
<i>Preface</i>	xi
<i>Abbreviations</i>	xiii
1. Introduction: Analogy in grammar <i>James P. Blevins and Juliette Blevins</i>	1
Part I. Typology and Complexity	
2. Principal parts and degrees of paradigmatic transparency <i>Raphael Finkel and Gregory Stump</i>	13
3. Parts and wholes: Implicative patterns in inflectional paradigms <i>Farrell Ackerman, James P. Blevins, and Robert Malouf</i>	54
4. Resolving pattern conflict: Variation and selection in phonology and morphology <i>Andrew Wedel</i>	83
Part II. Learning	
5. The relation between linguistic analogies and lexical categories <i>LouAnn Gerken, Rachel Wilson, Rebecca Gómez, and Erika Nurmsoo</i>	101
6. The role of analogy for compound words <i>Andrea Krott</i>	118
7. Morphological analogy: Only a beginning <i>John Goldsmith</i>	137
Part III. Modeling Analogy	
8. Expanding Analogical Modeling into a general theory of language prediction <i>Royal Skousen</i>	164
9. Modeling analogy as probabilistic grammar <i>Adam Albright</i>	185

10. Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation	214
<i>Petar Milin, Victor Kuperman, Aleksandar Kostić, and R. Harald Baayen</i>	
<i>References</i>	253
<i>Index</i>	273

Notes on Contributors

Farrell Ackerman works on a range of syntactic and morphological issues viewed from the perspective of lexicalism and construction-theoretic approaches, with a focus on Uralic languages. He has worked on complex predicates (Ackerman and Webelhuth, *A Theory of Predicates*, 1998), argument-encoding and mapping theories (Ackerman and Moore, *Proto-properties and grammatical encoding*, 2001), and the typology of relative clauses (Ackerman, Nikolaeva, and Malouf, *Descriptive Typology and Grammatical Theory*, forthcoming). Over the past few years he has been exploring word-based implicative models of morphology. He is Professor of Linguistics and Director of the Interdisciplinary Graduate Program in Human Development at the University of California, San Diego.

Adam Albright received his BA in linguistics from Cornell University in 1996 and his Ph.D. in linguistics from UCLA in 2002. He was a Faculty Fellow at UC Santa Cruz from 2002–4, and since then has been an Assistant Professor at the Massachusetts Institute of Technology. His research interests include phonology, morphology, and learnability, with an emphasis on using computational modeling and experimental techniques to investigate issues in phonological and morphological theory.

Harald Baayen is Professor of Linguistics at the University of Alberta, Edmonton. His research interests include lexical statistics in literary and linguistic corpus-based computing, general linguistics, morphological theory, and the psycholinguistics of morphological processing. He has published in, e.g., *Language*, *Linguistics*, *Folia Linguistica*, *Computational Linguistics*, *Yearbook of Morphology*, *Computers and the Humanities*, *Literary and Linguistic Computing*, *Journal of Quantitative Linguistics*, *Journal of Memory and Language*, *Journal of Experimental Psychology*, *Language and Cognitive Processes*, *Cognition*, *Brain and Language*, and the *Philosophical Transactions of the Royal Society* (Series A: Mathematical, Physical and Engineering Sciences) and has a new book, *Analyzing Linguistic Data: A Practical Introduction to Statistics using R* (Cambridge University Press, 2008).

James P. Blevins is Assistant Director of Research at the Research Centre for English and Applied Linguistics in the University of Cambridge. He received his Ph.D. in linguistics from the University of Massachusetts, Amherst, in

1990, and has taught at the University of Western Australia, the University of Texas, the University of Alberta, and the University of California, Berkeley. His research deals mainly with the description and analysis of morphological systems and syntactic constructions, with a particular emphasis on paradigm structure and discontinuous dependencies. Areal interests include Germanic, Balto-Finnic, Balto-Slavic, and Kartvelian.

Juliette Blevins is a Senior Scientist in the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig. She received her doctorate in linguistics from MIT in 1985, and then joined the Department of Linguistics at the University of Texas at Austin. Her research interests range from historical, descriptive, and typological studies, to theoretical analysis with a synthesis in her recent book *Evolutionary Phonology* (Cambridge University Press). Other interests include Oceanic languages, Australian Aboriginal languages, and Native American languages.

Raphael Finkel received a Ph.D. from Stanford University in 1976 in robotics and has been a professor of computer science at the University of Kentucky in Lexington since 1987. His research involves operating system administration, computational morphology, reliable data backup, and web-based organic chemistry homework. He was associated with the first work on quad trees, k-d trees, quotient networks, and the Roscoe/Arachne, Charlotte, and Unify operating systems. He helped develop DIB, has published over eighty articles, and has written two textbooks: *An Operating Systems Vade Mecum* (Prentice-Hall, 1988), and *Advanced Programming Language Design* (Benjamin-Cummings, 1996). He is also a co-author of *The Hacker's Dictionary* (Harper and Row, 1983).

LouAnn Gerken received her Ph.D. in experimental psychology from Columbia University in 1987. She is currently Professor of Psychology and Linguistics and Director of Cognitive Science at the University of Arizona. The focus of her research is behavioral indicators of the computational mechanism by which human infants and young children learn the structure of their native language. In addition, she works on issues of social justice in the academy.

John Goldsmith is a professor in the departments of Linguistics and Computer Science at the University of Chicago. His research interests include phonological theory, especially autosegmental phonology and harmonic phonology, implementations of linguistic theory using current tools of machine learning, and the history of linguistics and the other mind sciences. His work on automatic grammar induction has focused on the problems of inferring

morphological structure from raw data, and is available at <<http://linguistica.uchicago.edu/>>.

Rebecca Gómez obtained her Ph.D. in experimental psychology at New Mexico State University in 1995. She completed a postdoctoral position at the University of Arizona before taking her first faculty position at Johns Hopkins University in 1999. She returned to the University of Arizona in 2001 where she is presently an Associate Professor. She is known for her work on learning in cognitive development and child language acquisition.

Andrea Krott is a lecturer in the School of Psychology at the University of Birmingham. She holds an M.A. from the University of Trier, Germany, and a Ph.D. from the Radboud University Nijmegen, The Netherlands. Her research focuses on the processing and acquisition of word morphology. In particular, she investigates the role of analogy in the production and comprehension of noun-noun compound words in adult speakers across different languages and its role in preschool children's understanding of compound words. Her interests also include electrophysiological measures of language processing in adults.

Victor Kuperman is a Ph.D. candidate at Radboud University Nijmegen and a researcher at the Max Planck Institute for Psycholinguistics in Nijmegen. He is engaged in information-theoretical research of production and comprehension of morphologically complex words. His interests include morphological and sentence processing in visual and auditory domains; interaction of phonological and orthographic encodings in silent reading; sex differences in the uptake and processing of visual information, as revealed in eye movements; and statistical methods in language research.

Robert Malouf received his Ph.D. in linguistics from Stanford University in 1998. He has held positions at Stanford University, UC Berkeley, and the University of Groningen, and is currently an assistant professor in the computational linguistics program of the Department of Linguistics and Asian/Middle Eastern Languages at San Diego State University. His research focuses on constructional approaches to morphology and syntax, statistical natural language processing, and the application of text-mining techniques to research in theoretical linguistics.

Petar Milin is an assistant professor of research methodology and cognitive psychology at the Department of Psychology, University of Novi Sad, and holds a research position at the Laboratory for Experimental Psychology, University of Belgrade. He received his Ph.D. in psychology from the University of

Belgrade in 2004. His research interests include the psycholinguistics of morphological processing, probabilistic models of language processing, and lexical statistics. Together with **Aleksandar Kostić**, he has started a publishing project devoted to frequency dictionaries and electronic corpora of famous Serbian writers.

After being awarded a Ph.D. in Psychology from Yale University in 2006, **Erika Nurmsoo** worked as a postdoctoral research fellow at the University of Warwick in England. She is currently a research fellow at Bristol University, UK, studying children's cognitive development.

Royal Skousen is Professor of Linguistics and English Language at Brigham Young University. He has also taught at the University of Illinois, the University of Texas, the University of California at San Diego, and the University of Tampere in Finland. He is the author of *Analogical Modeling of Language* (1989) and *Analogy and Structure* (1992). More recently, he has developed a quantum mechanical approach to Analogical Modeling that uses quantum computing to avoid the exponential processing time inherent in sequential models of Analogical Modeling. Since 1988 he has been the editor of the critical text of the Book of Mormon.

Gregory Stump is Professor of Linguistics at the University of Kentucky. He is the author of *Inflectional Morphology* (Cambridge, 2001) and of numerous articles on morphological theory and typology. In recent years, his research has focused on the theoretical and typological significance of inflectional paradigms; he has asserted the need for an inferential-realizational theory of inflectional exponence and has demonstrated that the implicative relations among the parts of inflectional paradigms are an important domain of typological variation.

Andy Wedel began his academic life in molecular biology with a Ph.D. dissertation on bacterial control circuits, followed by postdoctoral work in in-vitro evolution of RNA enzymes. After discovering linguistics, he turned his experience in complex systems and evolution toward attempting to illuminate the ways in which feedback at lexical and community levels interacts with biases in variation to create and modify language patterns. His hobbies include Turkish and tall-grass prairie restoration.

After finishing her Ph.D. in 2002, **Rachel Wilson** went on to become a lawyer. She currently works at a non-profit organization representing victims of torture and persecution.

Preface

This volume grew out of a workshop, entitled “Analogy in Grammar: Form and Acquisition,” which took place on September 22 and 23, 2006 at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. This was the first workshop on analogy at the Institute, and quite possibly the first meeting in Leipzig dedicated to this topic since the late nineteenth century. At that time, analogy was a prominent concern of the *Junggrammatiker* (or Neogrammarians) at the University of Leipzig, who met often to discuss not only sound laws and their regularity, but exceptions to the regularity principle, most of which were explained with direct reference to analogy.

As organizers of the workshop, our goal was to gather researchers from a wide range of disciplines, in order to compare approaches to central questions about the form and acquisition of analogical generalizations in language, and to share results that have been obtained so far. The discussion was framed by a number of basic questions. What kinds of patterns do speakers select as the basis for analogical extension? What types of items are susceptible or resistant to analogical pressures? At what levels do analogical processes operate, and how do processes interact? What formal mechanisms are appropriate for modeling analogy? How does analogical modeling in cognitive psychology carry over to studies of language acquisition, language change, and language processing? The participants who addressed these questions included cognitive psychologists, developmental psychologists, psycholinguists, historical linguistics, descriptive linguists, phoneticians, phonologists, morphologists, syntacticians, computational linguists, and neurolinguists.

The bulk of the workshop was devoted to the presentation of original research. Papers and their authors were: “Paradigmatic heterogeneity” by Andrew Garrett; “Multi-level selection and the tension between phonological and morphological regularity” by Andrew Wedel; “Principal parts and degrees of paradigmatic transparency” by Rafael Finkel and Gregory Stump; “Analogy as exemplar resonance: Extension of a view of sensory memory to higher linguistic categories” by Keith Johnson; “Learning morphological patterns in language” by John Goldsmith; “The sound of syntax: Probabilities and structure in pronunciation variation” by Susanne Gahl; “Patterns of relatedness in complex morphological systems” by Farrell Ackerman and Robert Malouf; “Linguistic generalization by human infants” by LouAnn Gerken; “Banana shoes and bear tables: Children’s processing and interpretation of noun-noun

compounds” by Andrea Krott; “Analogy in the acquisition of constructions” by Mike Tomasello; “Acquisition of syntax by analogy: Computation of new utterances out of previous utterances” by Rens Bod; “Expanding analogical modeling into a general theory of language prediction” by Royal Skousen; “Analogical processes in learning grammar” by Dedre Gentner; “Modeling analogy as probabilistic grammar” by Adam Albright; and “Bits and pieces of an information-theoretical approach to inflectional paradigms” by Harald Baayen and Fermín Moscoso del Prado Martín. Formal commentary was provided by Colin Bannard on analogical modeling and computation, by Elena Lieven on analogy in language acquisition, and by Jim Blevins on modeling approaches. Not all of these authors were able to contribute to this volume directly, but the stimulating discussions which followed each talk, and continued over coffee and beer, can be felt throughout. We thank all the participants in the workshop for their enthusiastic participation, thought-provoking papers, and gracious handling of comments and questions.

The Analogy Workshop would not have been possible without the financial support of the Max Planck Society. We are grateful to Bernard Comrie, Head of the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology, for generously agreeing to host this workshop, and for hosting Jim as a visiting scholar at the Institute. Additional thanks are due to Mike Tomasello, Head of the Department of Developmental and Comparative Psychology, to administrative assistants Claudia Büchel, Julia Cissewski, Eike Lauterbach, Martin Müller, Claudia Schmidt, and Henriette Zeidler, for ensuring that everything ran smoothly, and to Claudio Tennie, for the memorable zoo tour.

Thanks to the support and enthusiasm of John Davey and Julia Steer at Oxford University Press, the workshop led seamlessly to this book project. As chapters were submitted, a small group of dedicated referees offered useful commentary. We are grateful for their assistance, and respectful of their wishes to remain anonymous. The contributors, in addition to offering original research, made great efforts to write with a general linguistics audience in mind, and to relate their work to that of others. For all of this, we are greatly appreciative. Finally, we would like to take this opportunity to thank each other for all that went into the preparation of this volume.

Abbreviations

1	1st person
2	2nd person
3	3rd person
abl	ablative
acc	accusative
AM	Analogical Modeling
AML	Analogical Modeling of Language
avg	average
CARIN	Competition Among Relations in Nominals
CE	cross entropy
conj	conjugation class
dat	dative
D	disyllabic verb stem diacritic
du	dual
ERP	event-related potential
exp	experiment
F	falling tone
feat	feature
FSA	finite state automaton
GCM	Generalized Context Model
gen	genitive
H	high tone
indic	indicative
iness	inessive
inf	infinitive
ins	instrumental
IPA	International Phonetic Alphabet
L	low tone
lexdec	lexical decision
loc	locative

LSA	Latent Semantic Analysis
m	mid tone
MDL	Minimum Description Length
MGL	Minimum Generalization Learner
N	noun
nom	nominative
part	partitive
PCFC	Paradigm Cell Filling Problem
perf	perfect
pl	plural
pres	present
pro	prolative
pros	prosecutive
QAM	Quantum Analogical Modeling
QM	Quantum Mechanics
RE	Relative Entropy
rel	relative
RT	reaction time
seg	segment
sg	singular
SPE	<i>The Sound Pattern of English</i>
subj	subjunctive
TiMBL	Tilburg Memory-Based Learner
V	verb
WP	Word and Paradigm
WPM	Word and Paradigm morphology

Introduction: Analogy in grammar

James P. Blevins and Juliette Blevins

1.1 Analogy: The core of human cognition

The human mind is an inveterate pattern-seeker. Once found, patterns are classified, related to other patterns, and used to predict yet further patterns and correlations. Although these tasks are performed automatically, they are far from trivial. The analogical reasoning that underlies them requires the discovery of structural similarities between perceptually dissimilar elements. Similarities may be highly abstract, involving functional and causal relationships. And while the recognition of analogical relations may seem like a passive process, it is in fact an aggressive process, driven by a search for predictability. A systematic structural similarity independent of perceptual similarity can be extended to yield novel inferences about the world.

There is mounting evidence from work in cognitive psychology that the talent for analogical reasoning constitutes the core of human cognition (Penn, Holyoak, and Povinelli 2008, and references cited therein), and that analogy may be a highly domain-independent cognitive process (Halford and Andrews 2007). Analogy is part of what allows humans to evaluate cause and effect, to come up with new solutions to old problems, to imagine the world other than the way it is, and to use words evocatively (Gentner, Holyoak, and Kokinov 2001). Other creatures create and use complex tools (Hunt and Gray 2004) and meta-tools (Taylor *et al.* 2007), recognize perceptual similarity and, after training, can perform better than chance on tests in which two objects must be judged as “same” or “different” (Premack 1983; Pepperberg 1987; Katz and Wright 2006). However, only humans, the symbolic species (Deacon 1997), effortlessly go beyond perceptual similarities, to recognize structural similarities that are independent of surface difference (Penn, Holyoak, and Povinelli 2008). Children as young as 1 or 2 years of age show evidence of perceptual analogies, and by the age of 4 or 5, they can understand that *bird* is to *nest*, as *dog* is to *doghouse*, using functional

analogies based on real-world knowledge (Goswami and Brown 1989, 1990; Goswami 2001).

As a central and pervasive property of human cognitive function and categorization, it is not surprising that analogy has been identified as a core component of linguistic competence from the earliest times to the present. In ancient Rome, Varro (116–127 BC) saw *analogia* as a central grammatical process (Law 2003), while ancient Arabic grammarians used the term *qiyaas* ‘measuring’ in a similar way: constructing a *qiyaas* involved “exploring an unknown configuration of data and trying to recognize in it a patterning already met and which, in other situations, lent itself to analysis” (Bohas, Guillaume, and Kouloughli 1990: 23). A thousand years later, analogy was central to one of the most important discoveries in linguistic history: the Neogrammarian insight that sound change was regular (Paul 1880/1920). Regular sound change, in contrast to analogy, was the foundation of the comparative method by which the world’s major language families were firmly established (Campbell and Poser 2008). To this day, regular sound change and analogy are introduced together to students of historical linguistics as the primary internal mechanisms of change (Hock 1991; Campbell 1998; Deutscher 2005), as research on the nature of analogical change continues (e.g. Lahiri 2000; Garrett 2008; Albright 2008). From its central role in historical linguistics, analogy became a cornerstone of analysis in the twentieth-century American descriptivist tradition (Whitney 1875/1979; Bloomfield 1933: 275; Sturtevant 1947: 96–109; Hockett 1966: 94) and, despite generative neglect, remains central to our understanding of synchronic grammars to this day (Anttila and Brewer 1977; Skousen 1989, 1992; Skousen, Lonsdale, and Parkinson 2002; Itkonen 2005; Kraska-Szlenk 2007).

The notion of analogy discussed above refers to a general cognitive process that transfers specific information or knowledge from one instance or domain (the analogue, base, or source) to another (the target). Sets of percepts, whether visual images, auditory signals, experiences, or dreams, are compared, and higher-order generalizations are extracted and carried over to new sets. This knowledge transfer is often schematized in terms of classical “proportional” or “four-part” analogies. In a proportional analogy, the relationship *R* between a pair of items *A*:*B* provides a basis for identifying an unknown item, given an item that matches *A* or *B*. Knowing *R* and knowing that *C* is similar to *A* permits one to identify *D* as the counterpart of *B*. The analogical deduction that “*C* is to *D*” as “*A* is to *B*” is standardly represented as in (1). The initial recognition of similarity and difference between percepts is the basis of analogy, but this is only the first step. Humans show great creativity in classifying different ways objects can be similar and different, and

in organizing these similarities and differences into complex schemata, which can then be extended to classify and understand new stimuli (Penn, Holyoak, and Povinelli 2008, and references cited therein).

(1) Four-part analogy: A is to B as C is to D

	A	:	B	==	C	:	D
a.	●	:	◐	==	■	:	■
b.	+ * +	:	* + *	==	XOX	:	OXO
c.	BIRD	:	NEST	==	DOG	:	DOGHOUSE

In (1a), one can look at the two circles ● : ◐ and establish a structural relationship between the two which is more general than the concrete circle and black and white shadings of its halves. The relationship could be extremely general: one figure is the reflection of the other. This structural relationship can then be recognized in other pairs, like the two squares, ■: ▣, and, in this general form, could be further extended to images without shading, like d:b, or Xxx:xxX. In (1b), there is a recognizable pattern ABA: BAB, where “A” and “B” can be replaced by any symbols, and where a more general statement of the pattern would allow reference to tones, melodies, or even conversational turn-taking. In (1c), where words in small capitals refer to concepts, the abstract structural relationship is a functional one relating an animal to its home or sleeping place. Again, the human mind is creative and flexible, and we can imagine the analogy extending to inanimate objects (CONTACT LENS : LENS CASE), musical traditions (JAZZ : NEW ORLEANS), or human emotions (ANGER : SPLEEN).

Before turning to the particular role that analogy plays in grammar, it is worth highlighting some general aspects of these relational patterns. First, although the analogues in (1) constitute paired objects, strings, and concepts, there is, in principle, no limit to how internally complex the analogue or base can be. We recognize human families, as well as language families, with mother tongues, daughter languages, and sister dialects. In language too, words can come in families, with complex kinship relations. These word families, often called **paradigms**, are a central locus of analogy in grammar. Inflectional paradigms can be extremely small, as in English {*dog, dogs*} or too large to list here, as in the approximately 1,000 inflected forms of a common Yupik (Eskimo) verb. The size of word families can also be highly item-specific within a language, as illustrated by the variation in the size of derivational paradigms in many languages, variation that is reflected in morphological family size effects (Bertram, Schreuder, and Baayen 2000; de Jong, Schreuder, and Baayen 2000). Although word-based analogies are often expressed as four-part analogies like those in (1), when large word-families are

involved, analogy may be much more complex. The nature of these complex analogical patterns is explored in several papers in this volume (Finkel and Stump; Ackerman, Blevins, and Malouf; and Milin *et al.*).

1.2 Analogy in grammar

In the domain of grammar, analogy is most strongly associated with language change (Anttila 1977; Hock 1991, 2003). Analogy is typically viewed as a process where one form of a language becomes more like another form due to an indirect association that is mediated by some higher-order generalization or pattern. While patterns can be observed across many linguistic categories, it is patterns between related words or word families that lead most often to analogical change. The short list of English singular and plural nouns in (2) exhibits a pattern that holds of the great majority of nouns in the language. Discounting compounds and derived forms, the families of these nouns are very small, consisting only of the two forms in (2).

(2) Some English singular and plural nouns

i. <u>Singular</u>	<u>Plural</u>	ii. <u>Singular</u>	<u>Plural</u>	iii. <u>Singular</u>	<u>Plural</u>
duck	ducks	kiss	kisses	baby	babies
cup	cups	dish	dishes	sister	sisters
sock	socks	fox	foxes	spoon	spoons
pot	pots	match	matches	apple	apples
chip	chips	lunch	lunches	bed	beds

Once a child has heard even a small set of nouns in the singular and plural, a pattern will start to emerge. The pattern relates a singular noun to a plural noun, where the plural noun is typically identical to the singular, except that it includes a predictable ending: /s/ after voiceless nonstrident sounds {p, t, k, f, θ} (2i); /əz/ after strident sounds {s, z, ʃ, ʒ, tʃ, dʒ} (2ii), and /z/ elsewhere (2iii). A proportional analogy like *sister:sisters = brother:X*, allows a child acquiring English to aggressively predict plurals not yet encountered on the basis of the singular form. Analogy yields *brothers*, the modern English plural, though similar analogical reasoning presumably led earlier to the replacement of *brethren* by *brothers*. Child language is full of analogical formations of this kind (*oxes, fishes, sheeps*, etc.) as well as others based on less robust patterns (e.g. *goose:geese = mongoose: mongeese*). The most salient examples are those that differ from adult forms, resulting in the strong association between analogy and language change.

However, there is a growing body of empirical evidence that linguistic change is continuous throughout the lifetime of an individual (Harrington, Palethorpe, and Watson 2000; Sankoff and Blondeau 2007, and references cited therein). Patterns of change suggest that linguistic knowledge is acquired incrementally, and that there is a feeding relationship between the production and perception of speech, which results in an ongoing process of grammar development (Pierrehumbert 2006; Wedel 2006, 2007). If this perspective is broadly correct, it suggests that the modern dichotomy between synchrony and diachrony is misconceived and that analogy is panchronic, and integral to the constantly evolving linguistic system of the individual. Recent simulations that use production/perception feedback loops have shown considerable promise in modeling the evolution of syntactic, morphological, phonological, and phonetic aspects of linguistic systems, and the success of these models is often enhanced by the introduction of analogy (see, e.g. Sproat 2008; Wedel this volume, and references cited therein.)

As suggested above, many of the most robust analogies in language involve word families as in (2), and can be referred to as **word-based analogy** or **morphological analogy**. In these cases, a recurrent **sound pattern and meaning** runs through a set of words, and forms the basis of the abstract pattern that newly heard words are associated with. In many cases, these can be stated as four-part analogies, but, as recognized as early as Paul (1880/1920), and further supported by Finkel and Stump (this volume), and Ackerman, Blevins, and Malouf (this volume), larger word sets may be necessary to discover patterns of predictability within complex inflectional systems. Furthermore, word families need not be limited to those defined by inflection or derivation. As shown by Krott (this volume), compounds define word families within which analogical formation is robust, and indeed the only explanation available for certain patterns.

There is evidence of word-based analogy in every language where analogical patterns have been investigated. The attraction of analogical patterns may be due in part to the fact that they impose a measure of order on the typically arbitrary sound–meaning correspondences in a language. But why should words play a distinguished role? In the cognitive psychology literature, it has been argued that the validity or strength of an analogy is partly determined by the number of distinct points at which one domain or entity can be aligned with another (Gentner 1983; Holyoak and Thagard 1989; Gentner and Markman 1997). This structural alignment will be very strong in word families, like the singulars and plurals in (2), since words can be aligned at phonetic, phonological, categorial, and inflectional feature points. In linguistic terms, the more shared features of different types a set of words

has, the more likely the set will be used as the basis of analogical modeling (Skousen 1989). Evidence for a minimal degree of structural alignment in word-based analogy is presented in Gerken *et al.* (this volume).

Because it is so widespread, word-based analogy has given rise to the greatest number of descriptive generalizations and theoretical proposals. At the descriptive level, the bulk of analogical changes are analyzed as instances of *extension* or *leveling*. Extension is the case where an alternating pattern is introduced to a historically nonalternating paradigm: e.g. English irregular *drive-drove* is extended in some dialects to *dive*, so that *dive-dived* > *dive-dove*. Under leveling, paradigmatic alternations are eliminated, as in the regularization of any historically strong verb, such as *cleave-cleaved* replacing the older *cleave-clove* in some varieties of English. More theoretical proposals attempt to define the most common directions of analogical change, taking into account phonological, morphological, syntactic, and semantic information. The best known of these are Kuryłowicz's laws of analogy (Kuryłowicz, 1945–9/1995) and Mańczak's tendencies in analogical change (Mańczak 1958). Both authors summarize recurrent aspects of word-based analogical change, from tendencies for transparent inflection to extend and replace synthetic forms, to generalizations governing which meanings are associated with old and new forms once analogy has taken place. However, as more morphological systems have been explored, few, if any, of these generalizations have survived. In their place, we see more general proposals. Deutscher (2001) divides internal word-based analogical change into "extension" and "reanalysis," in parallel with the typology of internal syntactic change (Harris and Campbell 1995). In a similar vein, Garrett (2008) suggests that pure leveling, in the sense outlined above, does not exist: instead, all cases of leveling are analyzed as extensions of an existing uniform paradigm on a nonuniform paradigm. Baayen *et al.* (2003*b*) demonstrates the importance of probabilistic knowledge in modeling morphological productivity, while Albright (2008) emphasizes an association between analogues and general informativeness.

Word-based analogies are by far the most widely recognized and carefully studied type, and their effects on language change are most salient. Nevertheless, analogy in grammar need not be limited to word-based comparisons, and cases involving phonetic, phonological, syntactic, and semantic alignment have also been proposed. In the domain of sound patterns, **phonetic analogy** is the case where a phonetically based variant of a particular segment is extended to another segment type or another context on the basis of phonetic similarity between segments or contexts (Bloomfield 1933: 366; Vennemann 1972; Steriade 2000; Yu 2007; Mielke 2008: 88–95). For example, in Tigrinya, velar stops /k/ and /g/ undergo spirantization to [x] and [ɣ]

respectively between vowels. In one dialect, spirantization has been extended to /b/ and /p/ as well, but not to /t/ or /d/. One analysis of this pattern is that the original velar spirantization is extended to labials, but not coronals, on the basis of analogy: labials and velars are phonetically similar, both being grave, with greater acoustic energy in the lower frequencies (Mielke 2008: 89–90). Though in some cases, alternative, purely phonetic, analyses are possible, and well supported (e.g. Barnes and Kavitskaya 2002, on French schwa deletion), it remains to be seen whether all cases can be dealt with in similar ways.

Direct sound–meaning or phonology–semantics alignments that are not mediated by the lexicon are usually characterized as systems of **sound symbolism** (Hinton, Nichols, and Ohala 1994). Conventional sound symbolism, where sound–meaning correspondences are highly language-specific, and to some extent arbitrary, provide the best examples of **phonological analogy**, especially where **phonaesthemes** are involved. Phonaesthemes are recurring sound–meaning pairs that cannot be construed as words or as morphemes, like English word-initial *gl-* in *glitter*, *glisten*, *glow*, *gleam*, *glint* which evokes light or vision (Firth 1930; Bloomfield 1933; Bergen 2004). Though they may arise by accidental convergence, the statistically significant distribution of sound–meaning pairs are interesting, in that they, like other patterns, are seized upon by language learners, forming the basis of productive analogies. As Bloomfield (1895: 409) observed colourfully: “Every word, in so far as it is semantically expressive, may establish, by hap-hazard favoritism, a union between its meaning and any of its sounds, and then send forth this sound (or sounds) upon predatory expeditions into domains where the sound is at first a stranger and parasite. . . .” In the case of English phonaesthemes, the psychological reality of the sound–meaning correspondence is evident in priming experiments (Bergen 2004), as well as in neologisms, where the correspondence is extended analogically (Magnus 2000). Looking for a new dishwashing powder? “Everything glistens with *Glist*.” or so an advertising slogan would have us believe. Direct sound meaning alignments need not be mediated by discrete phonological units. Words may have their own “gestalts,” or holistic patterns, and these may also be the basis of productive analogies (Hockett 1987).

Semantic analogies are usually classified as **metaphors**. In semantic analogies, relations between aspects of meaning of the analogue are mapped to those of the target (Gentner *et al.* 2001*b*). Though words are used to express semantic analogies, it is clear that, in some cases, words are merely vehicles for deeper conceptual alignments. The use of space to talk about time is a clear example: *a long illness; a short recovery; two weeks in advance; one month behind schedule*, etc. Cross-linguistically the metaphorical relationship

between space and time is asymmetrical: people talk about time in terms of space more often than they talk about space in terms of time (Lakoff and Johnson 1980; Alverson 1994). A range of psychophysical experiments supports a conceptual, nonlinguistic basis for this asymmetry: subjects take irrelevant spatial information into account when judging duration, but do not take special notice of irrelevant temporal information when judging space, providing evidence that semantic representations of time and space are inherently asymmetrical (Casasanto and Boroditsky 2007). Semantic analogies may also play a significant role in semantic change across time and space, and determine, in many cases, specific directions of grammaticalization, e.g. verbs > auxiliaries; verbs > adpositions; adpositions > case markers; ONE > indefinite markers; spacial adverbs > temporal adverbs (Traugott and Heine 1991; Heine 1993; Hopper and Traugott 2003).

Although highly intricate proposals have been advanced to account for syntactic knowledge, there is little counter-evidence to a very simple proposal. This classic model, which dominated language science until the rise of generative grammar, posits two basic mechanisms of human sentence production and comprehension (see, e.g. Sturtevant 1947:104–7). The first mechanism is memorization: people memorize utterances they have heard. These can range from very short phrases and simple sentences, to complex sentences, whole songs, poems or stories (Jackendoff 2002:152–4; 167–82). The second way in which people produce and understand phrases and sentences is by analogy with those they have memorized. In order to make use of **syntactic analogy**, a language learner must perform some segmentation of the utterance into smaller chunks (phrases or words) on the basis of sound/meaning correspondences. Based on this parsing, analogous bits or chunks of sentences can replace each other in different sentence frames (Tomasello 2003: 163–9). Two models that incorporate syntactic analogy have proved highly successful in accounting for syntactic acquisition and form. In language acquisition research, the “traceback” method analyzes dense corpora of child language in its natural context (Lieven *et al.* 2003; Dąbrowska and Lieven 2005). In the earliest stages of acquisition, one third of all children’s utterances are exact imitations of adult speech, while over 80 per cent of their speech is made up of exact copies of earlier utterances with only one analogically based operation (substitution, addition, deletion, insertion, or reordering). From utterances like *more milk*, *more juice*, the child is able to identify a frame “more N,” and extend it: *more jelly*, *more popsicle*, *more swimming*, etc. A similar perspective emerges from some of the models of construction grammar (Kay and Fillmore 1999; Croft 2001; Goldberg and Jackendoff 2005; Goldberg 2006), where syntactic productivity is viewed as the extension of learned constructions.

Constructions are the syntactic analogue of words: they typically embody arbitrary relations between form and meaning. The internal complexity of a construction, whose form may include phonological, morphological, syntactic, and pragmatic components, results in multiple anchor points for analogical extension.

A number of factors have contributed to the diminished role that analogy plays in generative accounts. The marginalization of morphology in general, and the neglect of complex inflectional systems in particular, shifted attention away from many of the patterns that traditional accounts had regarded in analogical terms. A primary focus on synchronic description likewise eliminated much of the traditional evidence for the influence of analogical pressures on the development of grammatical systems. A model of grammar that conceives of the mental lexicon as a largely redundancy-free collection of minimal units also lacks the word stock that provided the traditional base for analogical extensions of word-based patterns. In addition, the use of symbolic “rules” to provide a discrete description of a linguistic system imposes a strict separation between “data” and “program”. This departs from the more exemplar-based conception of approaches that treat analogy as the principal creative mechanism in language and recognize the probabilistic nature of linguistic generalizations (Bod, Hay, and Jannedy 2003; Gahl and Yu 2006). Hence, while Chomsky’s early remarks on grammar discovery echo some aspects of the descriptivist tradition (which retained a role for analogy), they also assume the notion of a “structural pattern” that corresponds to item-independent rules, not individual constructions or instances of any type of expression:

A primary motivation for this study is the remarkable ability of any speaker of a language to produce utterances which are new both to him and to other speakers, but which are immediately recognizable as sentences of the language. We would like to reconstruct this ability within linguistic theory by developing a method of analysis that will enable us to abstract from a corpus of sentences a certain structural pattern, and to construct, from the old materials, new sentences conforming to this pattern, just as the speaker does. (Chomsky 1955/1975: 131)

In later writings, Chomsky is dismissive of analogy on the few occasions that he mentions it at all (Itkonen 2005: 67–76), and his general position seems to be that “analogy is simply an inappropriate concept in the first place” (Chomsky 1986: 32). Work within the generative tradition has tended likewise to think of rules as the basis of broad generalizations, reserving analogy for local, lexically restricted patterns. A particularly clear and accessible

exposition of this perspective is *Words and Rules* (Pinker 1999). However, from a traditional perspective, a rule can be understood as a highly general analogy. There is no need for any qualitative difference between general and restricted analogies, and it is entirely plausible to assume that their differences reside solely in the specificity of the pattern that must be matched to sanction an analogical deduction. A number of psycholinguistic studies provide a measure of support for this more uniform view of grammatical devices by showing that there is no stable behavioral correlate of posited differences between irregular items (stored “words”) and productive formations (outputs of “rules”). Instead, different types of frequency information appear to be of central importance in conditioning variation in speakers’ responses in the lexical access and recognition tasks that are used to probe the structure of the mental lexicon (Stemberger and MacWhinney 1986; Hay and Baayen 2002, 2005; Baayen *et al.* 2003*b*). One further virtue of a unified notion of analogy that subsumes general and restricted cases is that it can account for the competition between candidate analogies in terms of the natural trade-off between the specificity of an analogical pattern and the number of encountered instances that match the pattern. It may even be possible to model or measure the attraction exerted by competing analogies given the advances in psycholinguistic methods for probing the structure of the mental lexicon (Milin *et al.*, this volume) and advances in techniques for modeling the effects of lexical neighborhoods (Wedel, this volume).

At this particular point in the development of the field of linguistics, it is useful to be reminded of the pivotal role that analogy has played in earlier grammatical models and to appreciate its renewed importance in the emerging quantitative and data-driven methodologies that feature in many of the papers in this volume. Nearly all grammatical traditions have regarded analogy as a central determinant of the form and evolution of linguistic subsystems, though it is only with the advent of better modeling techniques that it has become possible to investigate the psycholinguistic reality of analogical patterns and to represent and even measure the analogical pressures on a system. From this standpoint, it is perhaps the generative attitudes toward analogy that appear anomalous, a point that adds a further dimension to the reappraisal of generative approaches that is currently underway in phonology (Bybee 2001; J. Blevins 2004, 2006*b*; Mielke 2008), morphology (Anderson 2004; Deutscher 2005; J. P. Blevins 2006*b*); and syntax (Goldberg 2006; Matthews 2007; J. P. Blevins 2008). However one reconciles generative scepticism about analogy with more traditional perspectives, it would seem that this is an auspicious time to reconsider the role of analogy in grammar. In the chapters that follow, authors seek to understand better the ways in which

analogical reasoning, the core of human cognition, shapes the form and acquisition of linguistic knowledge.

1.3 Organization of this volume

The papers in this volume are organized thematically into three parts. The papers in each part address a group of related or overlapping issues, usually from slightly different or complementary perspectives.

The papers in Part 1 consider aspects of the organization of linguistic systems and the levels at which analogy operates in these systems. The central role attributed to analogy in morphological analysis is clear in the practice of matching principal parts against cells of exemplary paradigms to deduce unencountered forms. Yet although the deductions themselves can be represented by proportional analogies, many other aspects of this analysis remain imprecise, notably the criteria that guide the selection of principal parts. In Chapter 2, Finkel and Stump address this issue by proposing a typology of principal part systems, and by developing a notion of “paradigmatic transparency” that measures the degree of predictability between principal parts and paradigm cells. The information-theoretic approach outlined by Ackerman, Blevins, and Malouf in Chapter 3 offers a complementary perspective on this issue by representing implicational structure in terms of uncertainty reduction. In Chapter 4, Wedel sets out some of the ways that the organization of linguistic systems can evolve, reflecting different initial biases in a system or different ways of resolving conflicts between analogical pressures that operate at phonological and morphological levels.

The papers in Part 2 turn to the role that analogy plays in language learning, by humans but also by machines. In Chapter 5, Gerken *et al.* suggest that analogical reasoning about “secondary cues” accounts for the facilitatory effect that these cues apparently exert in the learning of lexical categories on the basis of paradigm-completion tasks. In Chapter 6, Krott reviews the pervasive influence of analogy on the form of compound structures in a range of languages. In Chapter 7, Goldsmith summarizes a body of research that has been devoted to building a general model of automatic morphological analysis and examines the contribution that analogy can make to the learning algorithm of this model.

Goldsmith’s paper provides a natural transition to the papers in Part 3, which take up the challenge of modeling analogy formally. In Chapter 8, Skousen offers a concise synopsis of the theory of Analogical Modeling, and presents analyses that motivate particular extensions of this theory.

In Chapter 9, Albright considers three restrictions on analogical inference that he argues can be attributed to limitations of context-sensitive rules. In the final chapter, Milin *et al.* return to issues concerning the organization of linguistic systems and present a range of studies that indicate the predictive value of information-theoretic measures, and also suggest the psychological relevance of traditional notions of paradigms and inflection classes.

Taken together, these papers reflect a resurgence of interest in traditional approaches to the representation and extension of grammatical patterns. It is hoped that collecting these papers together in the present volume will help to highlight significant points of contact across different domains and encourage further investigation of the role of analogy in language structure and use.

Principal parts and degrees of paradigmatic transparency

Raphael Finkel and Gregory Stump

2.1 Principal parts and inflectional paradigms

It is natural to suppose that in the case of many lexemes, language users store some of the forms in an inflectional paradigm and use these stored forms as a basis for deducing the other forms in that paradigm. Given that hypothesis, how much storage should one assume? At the maximal extreme, there could be full storage; this conclusion is not implausible for highly irregular paradigms or for paradigms whose forms are exceptionally frequent. At the minimal extreme, by contrast, there could be storage of only the minimum number of forms in a paradigm that are necessary for deducing all of the paradigm's remaining forms. Principal parts embody this notion of a minimal extreme. Postulating principal parts does not, of course, commit one to the assumption that speakers store a lexeme's principal parts and nothing more, only to the assumption that they are the minimum that could be stored if unstored forms are to be deduced from stored ones.

Principal parts have a long history of use in language pedagogy; generations of Latin students have learned that by memorizing a verb's four principal parts (those exemplified in Table 2.1), one can deduce all remaining forms in that verb's paradigm. But because principal parts are a distillation of the implicative relations that exist among the members of a lexeme's paradigm, they also reveal an important domain of typological variation in morphology.

In this paper, we use principal parts to identify a crucial dimension of this typological variation: that of PARADIGMATIC TRANSPARENCY—intuitively, the ease with which some cells in a paradigm can be deduced from other cells in that paradigm. We begin by distinguishing two types of principal-part analyses: static and dynamic (§2.2). Drawing upon principal-part analyses of the latter type, we develop a detailed account of paradigmatic transparency. For

TABLE 2.1 Principal parts of five Latin verbs

Conjugation	1st person singular present indicative active	Infinitive	1st person singular perfect indicative active	Perfect passive participle (neuter nominative singular)	Gloss
1st	laudō	laudāre	laudāvī	laudātum	‘praise’
2nd	moneō	monēre	monuī	monitum	‘advise’
3rd	dūcō	dūcere	dūxī	dūctum	‘lead’
3rd (-iō)	capiō	capere	cēpī	captum	‘take’
4th	audiō	audīre	audivī	auditum	‘hear’

concreteness, we exemplify our account by reference to the conjugational system of the Comaltepec Chinantec language (§2.3). Some of the conjugation classes in Comaltepec Chinantec give rise to maximally transparent paradigms; most others, however, deviate from maximal transparency in one or more ways (§2.4). We propose a formal measure of paradigm predictability to elucidate the degrees of such deviation (§2.5). The observable degrees of deviation from maximal transparency in both Comaltepec Chinantec and Fur turn out to be irreconcilable with the No-Blur Principle (Cameron-Faulkner and Carstairs-McCarthy 2000), according to which the affixes competing for the realization of a particular paradigmatic cell either uniquely identify a particular inflection class or serve as the default affixal realization of that cell (§2.6). At the same time, the proposed measure of paradigm predictability affords a precise account of cross-linguistic differences in paradigmatic transparency, as we demonstrate in a comparison of the conjugational systems of Comaltepec Chinantec and Fur (§2.7). We summarize our conclusions in §2.8.¹

Our work here contributes to the (by now quite vast) body of work demonstrating the typological and theoretical significance of inflectional paradigms in the structure of natural languages. Much of the research in this area has focused on the importance of paradigms for defining relations of inflectional exponence (e.g. Matthews 1972; Zwicky 1985; Anderson 1992; Stump 2001); other research, however, has drawn particular attention to the

¹ An earlier version of this paper was presented at the workshop on Analogy in Grammar: Form and Acquisition, September 22–3, 2006, at the Max Planck Institute for Evolutionary Anthropology, Leipzig. We thank several members of the audience at that event for their helpful comments; thanks, too, to two anonymous referees for several useful suggestions. Thanks finally to Eric Rowland for the dodecagon in Fig. 2.1.

significance of implicative relations among the cells in a paradigm (e.g. Wurzel 1989, J. P. Blevins 2006*b*, Finkel and Stump 2007). Our concerns here relate most directly to the latter sphere of interest.

2.2 Two conceptions of principal parts

Before proceeding, we must distinguish two importantly different conceptions of principal parts in natural language. (See Finkel and Stump 2007 for additional discussion of this distinction.)

2.2.1 *The static conception*

According to the *STATIC* conception of principal parts, the same sets of morphosyntactic properties identify the principal parts for every inflection class for lexemes in a given syntactic category. To illustrate, consider the hypothetical inflection-class system depicted in Table 2.2. In this table, there are four morphosyntactic property sets, represented as W, X, Y, and Z; there are six inflection classes, represented as Roman numerals I through VI; for each realization of a morphosyntactic property set within an inflection class, there is a particular exponent, and these exponents are represented as the letters a through o. We might represent this system of inflection classes with static principal parts as in Table 2.3. In this table, the shaded exponents represent the principal parts for each of the six inflection classes: The three shaded principal parts in each inflection class suffice to distinguish it from the other five inflection classes. A static system of principal parts for the set of inflection classes in Table 2.2 gives each lexeme belonging to the relevant syntactic category three principal parts: its realizations for the property sets W, X, and Y.

This static conception of principal parts is in fact the traditional one: The principal parts for Latin verbs in Table 2.1 are static, because they represent the same four morphosyntactic property sets from one inflection class to another.

TABLE 2.2 A hypothetical inflection-class system

	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o

TABLE 2.3 Static principal parts for the hypothetical system

	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o

(1) Sample static principal-part specifications:

Lexeme L belonging to inflection class I : L_a, L_e, L_i Lexeme M belonging to inflection class IV : M_c, M_g, M_j Lexeme N belonging to inflection class VI : N_d, N_h, N_l 2.2.2 *The dynamic conception*

According to the DYNAMIC conception of principal parts, principal parts are not necessarily parallel from one inflection class to another. The hypothetical inflection-class system in Table 2.2 admits the dynamic system of principal parts in Table 2.4. Each inflection class has only one shaded cell. If we observe that a lexeme has the exponent **a** in the form expressing the morphosyntactic property set W, we can deduce that it belongs to inflection class I; if we instead find that it has the exponent **b** in the realization of the property set W, we deduce that it belongs to inflection class II; if it exhibits the exponent **f** in the realization of property set X, we know that it belongs to inflection class III; and so forth. In a way, the dynamic conception of principal parts is more economical than the static because it allows each inflection class in this hypothetical example to have only a single principal part.

It's important to note, though, that under the dynamic conception of principal parts, the lexical specification of a lexeme's principal part must specify the morphosyntactic property set which that principal part realizes. Consider the slightly more complicated hypothetical system of inflection classes in Table 2.5. Here, the exponent **g** realizes the property set X in inflection class IV, but this same exponent **g** realizes the morphosyntactic property set Z in inflection class VII. In representing lexemes for this hypothetical system, it does not suffice simply to specify that a lexeme has a realization involving the exponent **g** as its principal part, because this fact fails to indicate whether that lexeme belongs to inflection class IV or inflection class VII. So lexical specifications of principal parts under the dynamic

TABLE 2.4 Dynamic principal parts for the hypothetical system

	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o

TABLE 2.5 Dynamic principal parts for a slightly larger system

	W	X	Y	Z
I	a	e	i	m
II	b	e	i	m
III	c	f	j	n
IV	c	g	j	n
V	d	h	k	o
VI	d	h	l	o
VII	c	e	j	g

conception are pairings of morphosyntactic property sets with realizations, as in (2). We refer to such pairings as *CELLS*.

(2) Sample dynamic principal-part specifications:

Lexeme L belonging to inflection class I : W:L_a

Lexeme M belonging to inflection class IV : X:M_g

Lexeme N belonging to inflection class VI : Y:N_l

Lexeme O belonging to inflection class VII : Z:O_g

The static and dynamic conceptions of principal parts differ in the answer they give to the question “What are a lexeme’s principal parts?” In the static approach, a lexeme’s principal parts are a list of words realizing a corresponding list of morphosyntactic property sets invariant across inflection classes; but under the dynamic approach, a lexeme’s principal parts are an unordered set of cells (pairings of realizations with morphosyntactic property sets).

For both the static and the dynamic conception of principal parts, the relation between a lexeme’s principal parts and its nonprincipal parts is fundamentally analogical in nature: if the principal parts of Lexeme₁ and

Lexeme₂ express the same morphosyntactic properties and are alike in form, then the nonprincipal parts of Lexeme₁ are analogous in form and content to those of Lexeme₂. For instance, given that the static principal parts of the Latin verb *amāre* ‘love’ (*amō, amāre, amāvī, amātum*) are parallel in form and content to those of *laudāre* ‘praise’ (cf. Table 2.1), their nonprincipal parts are entirely analogous; the same is true under a dynamic analysis, in which *amāre* and *laudāre* each have only a single principal part (the perfect passive participle).

Analogical relations have a range of implications for language acquisition and processing, not all of which are directly relevant to our discussion here. The principal-part analyses that we develop in this paper are based on a language’s complete system of inflection classes. Accordingly, our analyses are meant to account for fluent, adult speakers’ inference of analogical certainties:

- (3) If lexeme L has a particular set of principal parts, then it is a certainty that it has the associated set of nonprincipal parts.

Our central interest is in showing that the conditions licensing inferences of this sort vary in their complexity, both across a language’s paradigms and across languages. At the same time, we regard this work as providing the theoretical underpinnings necessary for a broader range of future investigations into the role of analogy in language. It is plausible to assume that language learners also make inferences such as (3) but that their inferences are defeasible by counter-evidence (which then necessitates a reconception of the implicative relations among paradigmatic cells). It is likewise plausible that inferences such as (3) have a role in online language processing, though we suspect that the extent to which this is true depends critically on a lexeme’s degree of paradigmatic transparency; for instance, a lexeme whose paradigm only requires a single dynamic principal part may give rise to such online inferences much more reliably than one whose paradigm requires five principal parts. We also suppose that inferences such as (3) could be successfully pressed into service in machine learning algorithms.

In this paper, we restrict our attention to dynamic principal parts as the basis for our account of paradigmatic transparency. Moreover, we generally restrict our attention to optimal sets of dynamic principal parts, where a set S of dynamic principal parts is OPTIMAL for inflection class J iff there is no valid set of dynamic principal parts for J whose cardinality is less than that of S. For example, although the set of dynamic principal parts specified in (4) is perfectly valid for deducing the exponence of the four lexemes listed, it is not optimal, since the smaller set of dynamic principal parts in (2) is also valid for deducing this exponence.

(4) Sample dynamic principal-part specifications:

Lexeme L belonging to inflection class I : W:L_a, X:L_eLexeme M belonging to inflection class IV : X:M_gLexeme N belonging to inflection class VI : Y:N_ILexeme O belonging to inflection class VII : Z:O_g

For concreteness, we develop this account with reference to the system of dynamic principal parts embodied by the system of conjugations in Comaltepec Chinantec (Oto-Manguean; Mexico).

2.3 Conjugation classes in Comaltepec Chinantec

We begin with an overview of the system of conjugation classes in Comaltepec Chinantec. Verbs in Comaltepec Chinantec inflect in two ways: through stem modulation and through the addition of affixes. The affixes include (i) aspectual prefixes expressing the progressive, the intentive, and the completive aspects and (ii) pronominal suffixes expressing the first-person singular, the first-person plural, the second-person singular, and the third person. (Verbs with a second-person plural subject lack any pronominal suffix.) The forms in Table 2.6 exemplify these affixes, whose use is essentially constant across all of the language's sixty-seven conjugations.

Patterns of stem modulation also serve to distinguish the three aspects as well as four person/number combinations (first-person singular, first-person plural, second person, and third person). In accordance with these patterns, a stem's final syllable may vary in tone (the seven possibilities being low [L], mid [M], high [H], and the combinations [LM], [MH], [LH], and [HL]); in stress (we leave controlled stress unmarked and mark ballistic stress with [']); in length (we leave short syllables unmarked and mark long syllables with [:]); in its capacity to trigger tone sandhi (we leave nontriggers unmarked and mark triggers as [\$]); and in the presence or absence of final glottality ("open"

TABLE 2.6 Inflectional paradigm of the verb 'play' (Conjugation P2B) in Comaltepec Chinantec

Aspect	1SG	1PL	2SG	3
Progressive	kó: ^L -R	ko: ^M -R?	ko: ^L -?	kó: ^L -r
Intentive	ni ^L -kó: ^{LH} -R	ni ^L -kó: ^H -R?	ni ^L -kó: ^H -?	ni ^L -kó: ^M -r
Completive	ka ^L -kó: ^M -R	ka ^L -kó: ^H -R?	ka ^L -ko: ^M -?	ka ^L -kó: ^L -r

R represents reduplication of a syllable-final segment; for details, see Pace 1990, Anderson *et al.* 1990.

(Source: Pace 1990: 42)

syllables—those lacking final glottality—we leave unmarked, and checked syllables—those exhibiting final glottality—we mark with [ʔ]).

A verb's membership in a particular conjugation class depends on (i) the particular pattern of stem modulation that its stem exhibits and (ii) the number of syllables in its stem. Stems are either monosyllabic or disyllabic; in the tables below, we represent disyllabic verb stems with the diacritic “d.”

Given these syllabic and prosodic differences among the conjugation classes, one can discern three broadly different groups of conjugations. In her description of Comaltepec Chinantec verb inflection, Pace (1990) calls these three groups Class A, Class B, and Class C verbs. Class A verbs, the largest such class, are represented in Table 2.7. Pace (1990: 43f.) distinguishes Class A verbs from verbs in the other classes by the following criteria:

In Class A verbs, first vs. nonfirst persons are distinguished in progressive aspect. Third person may also be distinguished. The three aspects have different inflectional patterns.

Within this broad characterization of Class A verbs, there is a range of variants; thus, there are thirty-five different conjugation classes represented among the Class A verbs in Table 2.7.²

Pace (1990: 46) uses the following criteria to distinguish the Class B verbs:

In Class B verbs, only third person is distinguished in noncompletive aspects. Like Class A verbs, the three aspects have different inflectional patterns.

As Table 2.8 shows, this broad characterization of Class B verbs subsumes nineteen different conjugations.

Finally, Pace (1990: 48) distinguishes Class C verbs by the following criteria.

In Class C verbs, third person is distinguished from nonthird. Aspect has different inflectional patterns in third person only.

² Here and further on, we represent an inflection class as a set of pairings of a particular morphosyntactic property set with an associated exponence; in any such pairing, the exponence may include affixes, prosodic markings, and phonological properties characteristic of the stems belonging to the inflection class at hand. This mode of representation reveals patterns that recur across two or more inflection classes (e.g. a pattern in which whatever exponence is associated with property set A is also associated with property set B); distinct inflection classes participating in patterns of this sort can then be grouped into superclasses, which can in turn be useful for expressing complex implicative relationships in the most general possible way. Here, we are not concerned with the task of superclassing, since a superclass does not, in itself, afford any economy in the number of principal parts required by the inflection classes that it subsumes.

Note that Conjugations P2A through P2G deviate from Pace's general description of Class A verbs insofar as their first-person singular stems are like their third-person stems in the progressive aspect. Note, too, that certain verbs (e.g. *tán^{LM}* ‘put into’) inflect as members of more than one conjugation; compare English verb forms such as *dreamed/dreamt*.

TABLE 2.7 Class A conjugations in Comaltepec Chinantec

Conj	Progressive				Intentive				Completive				Sample lexemes (cited by their 2nd person complete stem)
	1sg	1pl	2	3	1sg	1pl	2	3	1sg	1pl	2	3	
P1A	M	M	L	LM	MH	Hʔ	Hʔ	Mʔ	Mʔ	Hʔ	Lʔ	Mʔ	ká ^L ‘charge’
P1B	M	M	L	LM	MH	Hʔ	Hʔ	Mʔ	Mʔ	Hʔ	Hʔ	Lʔ	tá ^H ‘prune’
P1C	M	M	L	LM	MH	Hʔ	Hʔ	Mʔ	Mʔ	Hʔ	Lʔ	Lʔ	ʔt ^L ‘read’
P1D	M	M	L	LM	MH	Hʔ	Hʔ	Mʔ	Mʔ	Hʔ	Hʔ	LM	ŋt ^H ‘walk’, ná ^H ‘open’
P1E	M	M	L	LM	MH	Hʔ	Hʔ	Mʔ	Mʔ	Hʔ	LMʔ	LM	ná ^{LM} ‘open’
P2A	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	M	L:ʔ	kiu ^M ‘hit with fist’
P2B	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	M:	L:ʔ	ʔë: ^M ‘kick’
P2C	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	M:ʔ	L:ʔ	gi: ^M ‘tear’
P2D	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	L:	L:ʔ	ke: ^L ‘place’
P2E	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	LH:ʔ	L:ʔ	ʔt: ^{LH} ‘sell’
P2F	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	M:ʔ	Mʔ	tó: ^M ‘bake’
P2G	L:ʔ	M:	L:	L:ʔ	LH:ʔ	Hʔ	H:ʔ	Mʔ	Mʔ	Hʔ	LH:ʔ	Mʔ	tú: ^{LH} ‘pour out’
P3A	M\$ʔ’	M\$ʔ’	Lʔ’	Lʔ’	Hʔ’	Hʔ’	Hʔ	Lʔ’	Mʔ’	Hʔ	Lʔ	Lʔ’	təʔ ^L ‘apply’
P3B	M\$ʔ’	M\$ʔ’	Lʔ’	Lʔ’	Hʔ’	Hʔ’	Hʔ	Lʔ’	Mʔ’	Hʔ	Lʔ’	Lʔ’	húʔ ^L ‘cough’
P3C	M\$ʔ’	M\$ʔ’	Lʔ’	Lʔ’	Hʔ’	Hʔ’	Hʔ	Lʔ’	Mʔ’	Hʔ	LHʔ	Lʔ’	genʔ ^{LH} ‘swing’, koʔ ^{LH} ‘play with’, huënʔ ^{LH} ‘speak to’
P3D	M\$ʔ’	M\$ʔ’	Lʔ’	Lʔ’	Hʔ’	Hʔ’	Hʔ	Lʔ’	Mʔ’	Hʔ	LMʔ	Lʔ’	hŋanʔ ^{LM} ‘kill’
P3E	M\$ʔ’	M\$ʔ’	Lʔ’	Lʔ’	Hʔ’	Hʔ’	Hʔ	Lʔ’	Mʔ’	Hʔ	LMʔ’	Lʔ’	sénʔ ^{LM} ‘hold’, ʔnóʔ ^{LM} ‘look for’, tánʔ ^{LM} ‘put into’

(Continued)

TABLE 2.7 (Continued)

Conj	Progressive				Intentive				Completive				Sample lexemes (cited by their 2nd person completive stem)
	1sg	1pl	2	3	1sg	1pl	2	3	1sg	1pl	2	3	
P ₃ F	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	Mʔ'	Hʔ'	LMʔ'	Mʔ'	laʔ ^{LM} 'bathe'
P ₃ G	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	Mʔ'	Hʔ'	LMʔ'	Mʔ'	ʔnoʔ ^{LM} 'look for'
P ₃ H	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	M\$ʔ'	Hʔ'	Lʔ'	Lʔ'	béʔ ^L 'roll up'
P ₃ I	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	M\$ʔ'	Hʔ'	LMʔ'	Lʔ'	koʔ ^{LM} 'play with'
P ₃ J	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	M\$ʔ'	Hʔ'	LMʔ'	Lʔ'	hiʔ ^{LM} 'smell'
P ₃ K	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	Mʔ'	LHʔ'	LMʔ'	Lʔ'	huénʔ ^{LM} 'speak to'
P ₃ L	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Hʔ'	Hʔ'	Hʔ'	Lʔ'	Mʔ'	LHʔ'	LMʔ'	Mʔ'	tánʔ ^{LM} 'put into'
P ₄ A	M\$ʔ'	M\$ʔ'	LMʔ'	LMʔ'	Hʔ'	Hʔ'	Hʔ'	Mʔ'	Mʔ'	Hʔ'	Lʔ'	LMʔ'	ʔiénʔ ^L 'spray, wave'
P ₄ B	M\$ʔ'	M\$ʔ'	LMʔ'	LMʔ'	Hʔ'	Hʔ'	Hʔ'	Mʔ'	Mʔ'	Hʔ'	LHʔ'	LMʔ'	tènʔ ^{LH} 'drop'
P ₄ C	M\$ʔ'	M\$ʔ'	LMʔ'	LMʔ'	Hʔ'	Hʔ'	Hʔ'	Mʔ'	Mʔ'	Hʔ'	LMʔ'	LMʔ'	ciuʔ ^{LM} 'kiss, ʔienʔ ^{LM} 'spray, wave'
P ₁₂ A	M:	M:	L:	L:	Lʔ	Lʔ	Lʔ	Lʔ	Mʔ	Mʔ	Mʔ	L:	kuán ^M 'grow'
P ₁₂ B	M:	M:	L:	L:	Lʔ	Lʔ	Lʔ	Lʔ	Mʔ	Mʔ	M:ʔ	Mʔ	kó: ^M 'burn'
P ₁₂ C	M:	M:	L:	L:	Lʔ	Lʔ	Lʔ	Lʔ	Mʔ	Mʔ	L:	M:ʔ	iç: ^L 'swell'
P ₁₃ A	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Mʔ'	Mʔ'	Lʔ'	Lʔ'	Mʔ'	Mʔ'	Mʔ'	Mʔ'	róʔ ^M 'bear weight of'
P ₁₃ B	M\$ʔ'	M\$ʔ'	Lʔ'	Lʔ'	Mʔ'	Mʔ'	Lʔ'	Lʔ'	Mʔ'	Mʔ'	Lʔ'	Lʔ'	hínʔ ^L 'hiccough'
P ₁₆ A	DM\$ʔ	DMʔ	DLʔ	DLʔ'	DM\$ʔ	DMʔ	DLʔ	DLʔ'	DM\$ʔ	DMʔ	DLʔ	DLʔ'	hmi ^H ʔéʔ ^L 'defend'
P ₁₆ B	DM:	DM:	DHL:	DL'	DM:	DM:	DHL:	DL'	DM:	DM:	DHL:	DL'	hmi ^H kiu: ^{HL} 'toast, dry'
P ₁₆ C	DM\$ʔ	DM\$ʔ	DHLʔ	DLʔ'	DM\$ʔ	DM\$ʔ	DHLʔ	DLʔ'	DM\$ʔ	DM\$ʔ	DHLʔ	DLʔ'	hmi: ^L uiʔ ^{HL} 'smooth, plane'

(Source: Pace 1990: 43–6; 49–51)

TABLE 2.8 Class B conjugations in Comaltepec Chinantec

Conj	Progressive				Intentive				Completive				Examples
	1sg	1pl	2	3	1sg	1pl	2	3	1sg	1pl	2	3	
P5A	Lʔ	Lʔ	Lʔ	Lʔ	Hʔ	Hʔ	Hʔ	Lʔ	Lʔ	Hʔ	Lʔ	Lʔ	bá ^L ‘hit’
P5B	Lʔ	Lʔ	Lʔ	Lʔ	Hʔ	Hʔ	Hʔ	Lʔ	Lʔ	Hʔ	LMʔ	Lʔ	ʔá ^{LM} ‘wade across’
P6A	M:ʔ	M:ʔ	M:ʔ	M:ʔ	Hʔ	Hʔ	Hʔ	M:ʔ	M:ʔ	M:ʔ	Mʔ	M:ʔ	hli ^M ‘cover’
P6B	M:ʔ	M:ʔ	M:ʔ	M:ʔ	Hʔ	Hʔ	Hʔ	M:ʔ	M:ʔ	M:ʔ	M:ʔ	M:ʔ	hnú: ^M ‘rub against’
P6C	M:ʔ	M:ʔ	M:ʔ	M:ʔ	Hʔ	Hʔ	Hʔ	M:ʔ	M:ʔ	M:ʔ	LMʔ	M:ʔ	hín ^{LM} ‘scold’
P7A	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	M:	LMʔ	kuën ^M ‘give’
P7B	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	M:ʔ	LMʔ	ʔín ^M ‘pardon’, hnió:n ^M ‘drag’
P7C	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	L:	LMʔ	kuën ^L ‘give’
P7D	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	L:ʔ	LMʔ	hnió:n ^L ‘drag’
P7E	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	LH:ʔ	LMʔ	ʔŋi: ^{LH} ‘blow nose, spit’
P7F	LMʔ	LMʔ	LMʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	Mʔ	LH:ʔ	LMʔ	LMʔ	ʔín ^{LM} ‘pardon’
P14A	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	tá ^L ‘drop’
P14B	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Lʔ	Mʔ	Mʔ	Mʔ	Mʔ	ʔi ^M ‘enter’
P15A	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	Mʔ	ze ^M ‘go’
P15B	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	Mʔ’	huínʔ ^M ‘lazy’
PDBA	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	hmi ^L ʔi ^H ‘count’
PDBB	DHL:ʔ	DHL:ʔ	DHL:ʔ	DL:ʔ	DHL:ʔ	DHL:ʔ	DHL:ʔ	DL:ʔ	DHL:ʔ	DHL:ʔ	DHL:ʔ	DL:ʔ	hmi ^L gó: ^{HL} ‘deceive’
PDBC	DHʔ	DHʔ	DHʔ	DMʔ	DHʔ	DHʔ	DHʔ	DMʔ	DHʔ	DHʔ	DHʔ	DMʔ	hmi ^L ʔme ^H ‘sharpen’
PDBD	DMHʔ	DMHʔ	DMHʔ	DLHʔ	DMHʔ	DMHʔ	DMHʔ	DLHʔ	DMHʔ	DMHʔ	DMHʔ	DLHʔ	hmi ^L kʔ ^{MH} ‘help’

(Source: Pace 1990: 46–8; 50–1)

TABLE 2.9 Class C conjugations in Comaltepec Chinantec

Conj	Progressive				Intentive				Completive				Examples
	1sg	1pl	2	3	1sg	1pl	2	3	1sg	1pl	2	3	
P8A	M:	M:	M:	L:ʔ	M:	M:	M:	Mʔ	M:	M:	M:	L:ʔ	ʔme:nʔ ^M ‘hide’, na:n ^M ‘begin’
P8B	M:	M:	M:	L:ʔ	M:	M:	M:	Mʔ	M:	M:	L:	L:ʔ	na:n ^L ‘begin’
P9A	LH:ʔ	LH:ʔ	LH:ʔ	LMʔ	LH:ʔ	LH:ʔ	LH:ʔ	Mʔ	LH:ʔ	LH:ʔ	M:ʔ	LMʔ	kiá:n ^M ‘sweep’
P9B	LH:ʔ	LH:ʔ	LH:ʔ	LMʔ	LH:ʔ	LH:ʔ	LH:ʔ	Mʔ	LH:ʔ	LH:ʔ	L:ʔ	LMʔ	hí:n ^L ‘argue’
P9C	LH:ʔ	LH:ʔ	LH:ʔ	LMʔ	LH:ʔ	LH:ʔ	LH:ʔ	Mʔ	LH:ʔ	LH:ʔ	LH:ʔ	LMʔ	hú: ^{LH} ‘lie’, kiá:n ^{LH} ‘sweep’, hí:n ^{LH} ‘argue’
P10	LHʔ	LHʔ	LHʔ	LMʔ	LHʔ	LHʔ	LHʔ	Mʔ	LHʔ	LHʔ	LHʔ	LMʔ	hunʔ ^{LH} ‘squat down’
P11	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	LMʔ’	huínʔ ^{LM} ‘tire’
PCMA	M	M	M	M	M	M	M	M	M	M	M	M	ʔiu:n ^M ‘inside’
PCMB	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	LHʔ	niʔ ^{LH} ‘open out’
PCMC	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	LH:ʔ	ʔi:n ^{LH} ‘want’
PDCA	DM:	DM:	DM:	DM:	DM:	DM:	DM:	DM:	DM:	DM:	DM:	DM:	hmi: ^L ʔa:n ^M ‘hungry’
PDCB	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	DLMʔ’	hmi: ^L ʔínʔ ^{LM} ‘rest’
PDCC	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	DHʔ	hmi: ^L guānʔ ^H ‘bless’

(Source: Pace 1990: 48–51)

The different variant possibilities within Class C are represented in Table 2.9, with thirteen different conjugations.

We say that the paradigm P of a member of conjugation J is MAXIMALLY TRANSPARENT if each pairing of a property set with an exponent in P is unique across all conjugations to the paradigms of members of J . If lexeme L has a maximally transparent paradigm P , any cell in P can serve as L 's sole dynamic principal part.

Fig. 2.1 represents a maximally transparent paradigm having twelve cells. The numbers $\underline{1}$ through $\underline{12}$ in this diagram represent twelve different morphosyntactic property sets; the letters a through l represent the realizations of those twelve different property sets (so that each vertex in Fig. 2.1 is labeled as a cell); and each of the lines in this diagram represents a relation of bidirectional implication between two cells. In other words, the pairing of a realization with a morphosyntactic property set in every cell implies the pairing of a realization with a morphosyntactic property set in every other cell. If a language user has learned the implicative relations in which a maximally transparent paradigm P_1 participates, then upon learning that the paradigm P_2 of a newly encountered verbal lexeme has a cell analogous to a cell in P_1 , the language user can deduce every other cell in P_2 . Thus, transparency is associated with the ease with which some of the cells in a paradigm can be deduced from other cells in the same paradigm.

In the inflection of Comaltepec Chinantec verbs, there are (as in Fig. 2.1) twelve different morphosyntactic property sets. In the remainder of the paper,

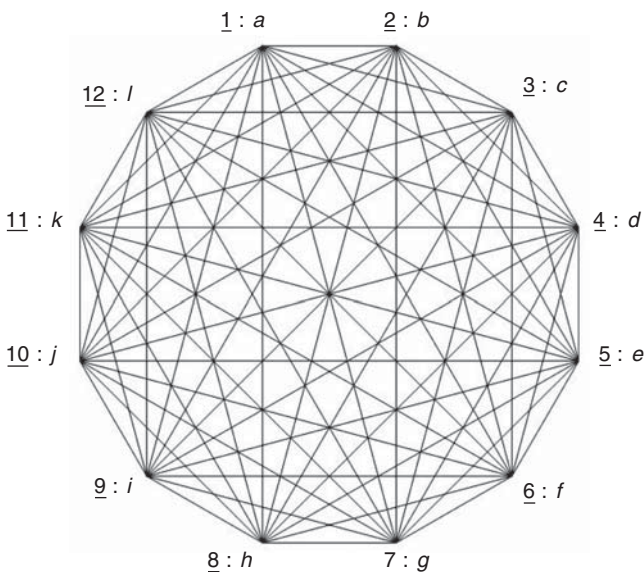


FIGURE 2.1 A maximally transparent paradigm with twelve cells

TABLE 2.10 Abbreviations for the twelve property sets realized by Comaltepec Chinantec verb forms

Abbreviation	Property set
<u>1</u>	1sg
<u>2</u>	1pl
<u>3</u>	2
<u>4</u>	3
<u>5</u>	1sg
<u>6</u>	1pl
<u>7</u>	2
<u>8</u>	3
<u>9</u>	1sg
<u>10</u>	1pl
<u>11</u>	2
<u>12</u>	3

we number these 1 through 12; the significance of these twelve numerals is given in Table 2.10.

The question now arises whether there are any maximally transparent paradigms in Comaltepec Chinantec. That is, are there conjugation classes whose paradigms could be represented as in Fig. 2.1? The answer is yes; in fact there are four such conjugations. One of these is Conjugation PDBB in Table 2.8, whose twelve alternative principal-part analyses are given in Table 2.11.

In Table 2.11, the cells in a lexeme's paradigm are given on the horizontal axis (where 1 through 12 represent the twelve morphosyntactic property sets corresponding to a verbal paradigm's twelve cells), and the different possible principal-part analyses are given on the vertical axis. Thus, each row represents a distinct principal-part analysis, and within a given row, the numeral *n* (any of the numerals 1 through 12) represents the morphosyntactic property set of the sole principal part in the principal-part analysis represented by that row. If a principal part *P* is listed in the column headed by a property set M in a given analysis, the realization of M is deducible from *P* in that analysis.

In Conjugation PDBB, any one of the cells in a lexeme's paradigm can be used as that lexeme's sole principal part—can be used, in other words, to deduce the realization of every one of the remaining eleven cells in the paradigm. This fact arises because each of the exponents of property sets 1 through 12 in Conjugation PDBB is unique to that conjugation. Table 2.12 lists the exponents of cells 1 through 12 in Conjugation PDBB; a comparison of these exponents with those given earlier in Tables 2.7-9 reveals that in each one of the twelve cells in the paradigm of a lexeme belonging to this conjugation, the exponence is absolutely distinctive of this conjugation.

TABLE 2.11 The twelve alternative optimal principal-part analyses for Conjugation PDBB in Comaltepec Chinantec

Alternative principal-part analyses	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9
10	10	10	10	10	10	10	10	10	10	10	10	10
11	11	11	11	11	11	11	11	11	11	11	11	11
12	12	12	12	12	12	12	12	12	12	12	12	12

TABLE 2.12 The exponence of property sets 1–12 in Conjugation PDBB in Comaltepec Chinantec

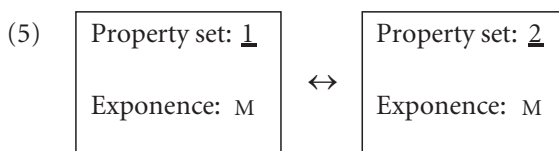
Morphosyntactic property sets												Example
<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	
DHL'	DHL'	DHL'	DL'	DHL'	DHL'	DHL'	DL'	DHL'	DHL'	DHL'	DL'	hmi ^L gô. ^{HL} 'deceive'

Only four conjugations have this property of maximal transparency in Comaltepec Chinantec; that is, most conjugations in this language deviate from maximal transparency. We now consider the consequences of this deviation.

2.4 Deviations from maximal transparency in Comaltepec Chinantec verb paradigms

Although every cell is fully informative in the paradigm of a verb belonging to Conjugation PDBB (and can therefore potentially serve as that verb's sole principal part), this full informativeness is comparatively rare. In the paradigms of most verbs, many cells are to some extent uninformative; that is, they have either a limited capacity or no capacity to serve as optimal principal parts. The system of conjugations in Comaltepec Chinantec exhibits various means of compensating for this less-than-full informativeness of certain cells.

Consider first a case in which a particular cell in a verb's paradigm uniquely determines only one other cell. In the paradigm of a verb belonging to Conjugation P1A, the cell containing the realization of property set 1 (whose exponence is tone M with controlled stress) uniquely determines the cell containing the realization of property set 2 (which has the same exponence), since no matter what the conjugation, the implicative relation in (5) holds true in Comaltepec Chinantec.



Even so, the cell containing the realization of property set 1 doesn't uniquely determine any of the remaining ten cells in the paradigm of a verb belonging to Conjugation P1A. In order to deduce the latter cells, the cell associated with property set 12 must be appealed to—either by itself or in addition to the cell associated with property set 1, as in Table 2.13. (In this table and those below, if a pair *P*, *Q* of principal parts is listed in the column headed by a property set M, the realization of M can only be deduced by simultaneous reference to *P* and *Q*.)

As Table 2.13 shows, the uninformative nature of the cell containing the realization of property set 1 in Conjugation P1A makes it necessary to deduce certain cells by simultaneous reference to two principal parts. The cells containing the realizations of property sets 1 and 12 are not, however, the only viable set of principal parts for a verb of this conjugation; another possibility is the set of cells containing the realizations of property sets 3 and 12, as in Table 2.14. Like the analysis in Table 2.13, the analysis in Table 2.14 requires two principal parts for this conjugation; although both analyses are optimal, the latter analysis might be preferred on the grounds that it makes lesser use of simultaneous reference to both principal parts in deducing the various nonprincipal parts.

As Table 2.14 shows, the number of nonprincipal parts that must be deduced by simultaneous reference to both principal parts can be minimized to one in Conjugation P1A. This sort of minimization isn't always possible, however. For

TABLE 2.13 A representative optimal principal-part analysis for Conjugation P1A in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>1,12</i>	<i>1</i>	<i>1</i>	<i>1,12</i>	<i>1,12</i>	<i>1,12</i>	<i>1,12</i>	<i>1,12</i>	<i>1,12</i>	<i>12</i>	<i>1,12</i>	<i>1,12</i>	<i>12</i>

TABLE 2.14 A representative optimal principal-part analysis for Conjugation P1A in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
3,12	3	3	3	3	3	3	3	3	3	3	3,12	12

TABLE 2.15 The sole optimal principal-part analysis for Conjugation P2C in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
11,12	11,12	12	11,12	12	11,12	11,12	11,12	12	11,12	11,12	11	12

TABLE 2.16 The sole optimal principal-part analysis for Conjugation P1B in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
11,12	11	11	11	11	11	11	11	11	11	11	11	12

instance, in the only optimal principal-part analysis for Conjugation P2C, seven of ten nonprincipal parts are deducible only by simultaneous reference to both principal parts; this analysis is given in Table 2.15.

Even so, the need to postulate two principal parts doesn't always entail a need for simultaneous reference to both of the principal parts in deducing one or another nonprincipal part. Consider, for example, Conjugation P1B, whose sole optimal principal-part analysis is given in Table 2.16. In the paradigm of a verb belonging to this conjugation, two principal-part specifications are necessary in order to deduce all of the remaining cells in the paradigm. In the only optimal analysis, the exponents of property sets 1 through 10 can all be deduced from the exponent of property set 11; the cell containing the realization of property set 11 is therefore one of the two principal parts of a verb belonging to this conjugation. The exponent of property set 12, however, cannot be deduced from that of property set 11, nor from that of any of the other property sets, and so must be independently specified. So here we have a conjugation that must have two principal parts, but no cell of which must be deduced by simultaneous reference to both principal parts.

TABLE 2.17 A representative optimal principal-part analysis for Conjugation P2F in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>1,11,12</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>1</i>	<i>11</i>	<i>12</i>

TABLE 2.18 The sole optimal principal-part analysis for Conjugation P3E in Comaltepec Chinantec

Principal parts	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>9,10,11,12</i>	<i>10</i>	<i>10</i>	<i>12</i>	<i>12</i>	<i>10</i>	<i>10</i>	<i>10</i>	<i>12</i>	<i>9</i>	<i>10</i>	<i>11</i>	<i>12</i>

Another example of the same type is Conjugation P2F, which involves three principal parts. In the representative principal-part analysis proposed in Table 2.17, the principal parts of a verb belonging to Conjugation P2F are the realizations associated with property sets 1, 11, and 12.

In Conjugation P3E, four principal parts are necessary. Table 2.18 represents the sole optimal principal-part analysis for this conjugation: The principal parts are the realizations of property sets 9, 10, 11, and 12.

In the deviations from maximal transparency that we have considered so far, the unformativeness of certain realizations has forced us to postulate two or more principal parts; some but not all of these analyses involve deducing certain nonprincipal parts by simultaneous reference to more than one principal part. But unformativeness needn't always lead to the postulation of more than one principal part. In some instances, it simply imposes limits on the range of alternative analyses.

For instance, a single principal part can be postulated for a verb belonging to Conjugation P16B, but there are only six cells in the paradigm of such a verb that can possibly serve as this sole principal part, namely the cells associated with property sets 3, 4, 7, 8, 11, and 12. The realizations in such a verb's paradigm can be deduced from any one of these cells but not from any other. Thus, a verb belonging to Conjugation P16B has the six alternative principal-part analyses represented in Table 2.19, but the realizations of property sets 1, 2, 5, 6, 9, and 10 in its paradigm are uninformative in any optimal principal-part analysis.

Conjugation P12A exhibits an even more severe restriction on the range of alternative analyses. In the paradigm of a verb belonging to this conjugation,

TABLE 2.19 The six optimal principal-part analyses for Conjugation P16B in Comaltepec Chinantec

Alternative principal-part analyses	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4
7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8
11	11	11	11	11	11	11	11	11	11	11	11	11
12	12	12	12	12	12	12	12	12	12	12	12	12

TABLE 2.20 The sole optimal principal-part analysis for Conjugation P12A in Comaltepec Chinantec

Principal part	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
12	12	12	12	12	12	12	12	12	12	12	12	12

only one cell (namely the cell associated with property set 12) can serve as the verb's sole principal part; the other eleven cannot. That is, the remaining realizations in such a verb's paradigm can be deduced from the cell containing the realization of property set 12 but not from any other cell. Thus, Conjugation P12A has the sole optimal principal-part analysis in Table 2.20; in this respect, it contrasts starkly with Conjugation PDBB (Table 2.11), any one of whose twelve cells may serve as its sole principal part.

The examples presented here show that the uninformative-ness of one or more cells in a lexeme's paradigm may have either or both of two effects on the principal-part analysis of that lexeme: (i) it may necessitate the postulation of more than one principal part for that lexeme; and (ii) it may limit the number of alternative optimal principal-part analyses to which that lexeme is subject. These effects therefore imply two practical criteria for paradigmatic transparency:

- (6) Two practical criteria for paradigmatic transparency
 - All else being equal,
 - a. fewer dynamic principal parts needed to deduce a lexeme's paradigm in an optimal analysis implies greater transparency of that paradigm;

- b. more alternative optimal principal-part analyses of a lexeme's paradigm implies greater transparency of that paradigm.

The transparency of Comaltepec Chinantec paradigms varies widely according to both of the criteria in (6), as we now show. Consider first Table 2.21, which relates to criterion (6a). As Table 2.21 shows, many of the conjugations involve paradigms that can be deduced from a single dynamic principal part; even more, however, have paradigms requiring two dynamic principal parts, and some require as many as three or even four dynamic principal parts. Thus, the successive rows in Table 2.21 represent decreasing levels of transparency according to criterion (6a).

Conjugations whose optimal analysis requires the same number of principal parts may nevertheless vary in the extent to which they require simultaneous reference to more than one principal part in deducing a cell's realization. Table 2.22 shows the average number of principal parts needed to deduce a cell's realization in each conjugation in Comaltepec Chinantec. The conjugation classes in rows A through D house verbs each of whose realizations can always be deduced by reference to a single principal part; those in the succeeding rows house verbs whose realizations must—to a progressively greater degree—be deduced through simultaneous reference to more than one principal part.

Table 2.23 relates to criterion (6b). Here the different conjugations are arranged according to the number of optimal principal-part analyses that

TABLE 2.21 Numbers of dynamic principal parts for Comaltepec Chinantec conjugation classes

Comaltepec Chinantec conjugation classes	Number of dynamic principal parts needed to identify a particular inflection class
P3A, P10, P11, P12A, P14A, P15A, P15B, P16A, P16B, P16C, PCMA, PCMB, PCMC, PDBA, PDBB, PDBC, PDBD, PDCA, PDCB, PDCC	1
P1A, P1B, P1D, P1E, P2A, P2B, P2C, P2D, P2E, P2G, P3C, P3F, P3H, P3I, P3J, P3K, P3L, P4A, P4B, P4C, P5A, P5B, P6A, P6B, P6C, P7A, P7B, P7C, P7D, P7E, P7F, P8A, P8B, P9A, P9B, P12B, P12C, P13A, P13B, P14B	2
P1C, P2F, P3D, P3G, P9C	3
P3B, P3E	4

TABLE 2.22 Average number of principal parts needed to identify a cell in Comaltepec Chinantec

Comaltepec Chinantec conjugation classes	Number of dynamic principal parts needed to deduce a lexeme's paradigm	Average number of principal parts needed to deduce a cell in a lexeme's paradigm
A. P _{3A} , P ₁₀ , P ₁₁ , P _{12A} , P _{14A} , P _{15A} , P _{15B} , P _{16A} , P _{16B} , P _{16C} , PCMA, PCMB, PCMC, PDBA, PDBB, PDBC, PDBD, PDCA, PDCB, PDCC	1	1.00
B. P _{1B} , P _{1D} , P _{1E} , P _{3H} , P _{3I} , P _{3J} , P _{3K} , P _{3L} , P _{4A} , P _{4B} , P _{4C} , P _{6A} , P _{6B} , P _{6C} , P _{7A} , P _{7B} , P _{7C} , P _{7D} , P _{7E} , P _{7F} , P _{8A} , P _{8B} , P _{9B} , P _{13A}	2	1.00
C. P _{1C} , P _{2F} , P _{3D} , P _{9C}	3	1.00
D. P _{3B} , P _{3E}	4	1.00
E. P _{1A} , P _{2A} , P _{2B} , P _{2D} , P _{12B} , P _{12C} , P _{13B}	2	1.08
F. P _{3F} , P _{9A}	2	1.25
G. P _{3G}	3	1.25
H. P _{5B}	2	1.33
I. P _{3C} , P _{14B}	2	1.42
J. P _{2C} , P _{2E}	2	1.58
K. P _{5A}	2	1.67
L. P _{2G}	2	1.75

they afford. The conjugation allowing the largest number of optimal principal-part analyses is P_{9C}, which allows twenty optimal analyses; but succeeding rows show conjugations allowing fewer analyses, with the bottom rows showing conjugations allowing only a single optimal principal-part analysis. Thus, by criterion (6b), Table 2.23 lists conjugation classes in decreasing order of paradigmatic transparency.

The application of criterion (6b) is complicated, however, by the fact that a paradigm is open to more alternative principal-part analyses the more principal parts it has. Thus, (6b) should be interpreted as meaning that the larger the number of principal-part analyses a conjugation has, the more transparent its paradigms are in comparison with those of other conjugations having the same number of principal parts. Where lexeme *L* has *k* principal parts and *n* is the number of morphosyntactic property sets for

TABLE 2.23 Numbers of optimal principal-part analyses for Comaltepec Chinantec conjugations

Conjugation	Number of principal parts	Number of optimal principal-part analyses
P9C	3	20
P12C	2	17
P14B	2	16
P3B	4	16
PDBB, PDBD, PDCB, PDCC	1	12
P11	1	11
P2A, P6A	2	11
P6B	2	10
PCMA	1	9
P7A, P7C, P7E, P12B	2	9
P6C, P13B	2	8
P15A	1	7
P1A, P5B, P7D, P8A, P9B, P13A	2	7
P1C, P2F	3	7
P16A, P16B, P16C, PDCA	1	6
P1E, P2B, P7B, P7E, P8B, P9A	2	6
P15B	1	5
P2D	2	5
P3C, P4A, P4C	2	4
PCMB, PCMC, PDBA, PDBC	1	3
P4B	2	2
P3A, P10, P12A, P14A	1	1
P1B, P1D, P2C, P2E, P3E, P3H, P3I, P3J, P3K, P3L, P5A	2	1
P3D, P3G	3	1
P3E	4	1

which L inflects, the largest possible number of optimal principal-part analyses for L is the binomial coefficient of n and k , i.e. $n!/(k!(n-k)!)$. The maximum possible number of optimal principal-part analyses for a Comaltepec Chinantec verb varies according to the number of principal parts it has, as in Table 2.24. Although the paradigm of a lexeme belonging to Conjugation P9C has the twenty alternative optimal principal-part analyses in Table 2.25, this paradigm is not all that transparent, since it has three principal parts, and is therefore far below the ceiling of 220 optimal analyses that a lexeme with three principal parts could imaginably have.

TABLE 2.24 Maximum possible number of optimal principal-part analyses for Comaltepec Chinantec verbs

Number (<i>k</i>) of principal parts	Maximum possible number $12!/(k!(12-k)!)$ of optimal principal-part analyses
1	12
2	66
3	220
4	495

TABLE 2.25 The twenty alternative optimal principal-part analyses for Conjugation P9C in Comaltepec Chinantec

Alternative principal- part analyses	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>1,4,11</i>	1	1	1	4	1	1	1	4	1	1	11	4
<i>1,8,11</i>	1	1	1	1,8	1	1	1	8	1	1	11	1,8
<i>1,11,12</i>	1	1	1	12	1	1	1	12	1	1	11	12
<i>2,4,11</i>	2	2	2	4	2	2	2	4	2	2	11	4
<i>2,8,11</i>	2	2	2	2,8	2	2	2	8	2	2	11	2,8
<i>2,11,12</i>	2	2	2	12	2	2	2	12	2	2	11	12
<i>3,4,11</i>	3	3	3	4	3	3	3	4	3	3	11	4
<i>3,8,11</i>	3	3	3	3,8	3	3	3	8	3	3	11	3,8
<i>3,11,12</i>	3	3	3	12	3	3	3	12	3	3	11	12
<i>4,5,11</i>	4,5	4,5	4,5	4	5	4,5	4,5	4	4,5	4	11	4
<i>4,6,11</i>	6	6	6	4	6	6	6	4	6	4	11	4
<i>4,7,11</i>	7	7	7	4	7	7	7	4	7	4	11	4
<i>4,9,11</i>	9	9	9	4	9	9	9	4	9	4	11	4
<i>5,11,12</i>	5,12	5,12	5,12	12	5	5,12	5,12	12	5,12	12	11	12
<i>6,8,11</i>	6	6	6	6,8	6	6	6	8	6	6	11	6,8
<i>6,11,12</i>	6	6	6	12	6	6	6	12	6	6	11	12
<i>7,8,11</i>	7	7	7	7,8	7	7	7	8	7	7	11	7,8
<i>7,11,12</i>	7	7	7	12	7	7	7	12	7	7	11	12
<i>8,9,11</i>	9	9	9	8,9	9	9	9	8	9	9	11	8,9
<i>9,11,12</i>	9	9	9	12	9	9	9	12	9	9	11	12

By the criteria in (6), Comaltepec Chinantec verb conjugations exhibit widely varying degrees of paradigmatic transparency. At the high extreme, that of total paradigmatic transparency, are the conjugations in (7a): lexemes in these conjugations exhibit only a single principal part and allow the

maximum number of alternative optimal principal-part analyses. At the opposite extreme is the conjugation in (7b): lexemes in Conjugation P_{3E} have four principal parts and allow only a single optimal principal-part analysis. Between these extremes, other lexemes exhibit a range of intermediate degrees of paradigmatic transparency.

- (7) Extreme degrees of paradigmatic transparency in Comaltepec Chinantec
 a. High: PDBB, PDBD, PDCB, PDCC b. Low: P_{3E}

2.5 A measure of paradigmatic transparency

Although the practical criteria in (6) are useful for distinguishing degrees of paradigmatic transparency, we would like to give more explicit content to the notion of paradigmatic transparency than these criteria allow. We therefore propose a precise measure of paradigmatic transparency; we call this measure PARADIGM PREDICTABILITY. The fundamental idea underlying this proposed measure is that where (i) M is the set of morphosyntactic property sets associated with the cells in the paradigm P_L of some lexeme L and (ii) M' is the set $\{N: N \subseteq M \text{ and the exponence in } P_L \text{ of the morphosyntactic property sets belonging to } N \text{ suffices to determine the exponence in } P_L \text{ of every morphosyntactic property set belonging to } M\}$, L 's paradigm predictability pp_L is calculated as in (8). In effect, this measure calculates the fraction of the members of M 's power set $\mathcal{P}(M)$ that are viable (though not necessarily optimal) sets of dynamic principal parts for L .

$$(8) \quad pp_L = \frac{|M'|}{|\mathcal{P}(M)|}$$

We refine this measure of paradigm predictability in two ways. First, the set M sometimes contains multiple morphosyntactic property sets whose exponence is the same across all inflection classes. We propose to eliminate all but one of these sets from M for purposes of calculating paradigm predictability. To understand why, consider the two hypothetical inflection-class systems in (9), in which i through iv represent inflection classes; s_1 through s_3 represent morphosyntactic property sets; and a through c represent inflectional exponents.

(9) System (9a)	System (9b)
$\underline{s_1 \quad s_2 \quad s_3}$	$\underline{s_1 \quad s_2}$
I a b b	I a b
II a c c	II a c
III b b b	III b b
IV c c c	IV c c

If paradigm predictability is calculated as in (8), then lexemes belonging to inflection class *I* in system (9a) have greater paradigm predictability than lexemes belonging to inflection class *I* in system (9b): the former have a predictability of $3/8$ ($M = \{s_1, s_2, s_3\}$, M' has three members $\{s_1, s_2\}$, $\{s_1, s_3\}$, $\{s_2, s_3\}$, and $\mathcal{P}(M)$ has eight), while the latter have a predictability of $1/4$ ($M = \{s_1, s_2\}$, M' has one member $\{s_1, s_2\}$, and $\mathcal{P}(M)$ has four). We prefer to think of lexemes in these systems as having the same predictability, namely $1/4$. To accommodate this preference, we let M_- be a maximal subset of M such that no two of members of M_- are identical in their exponence across all conjugations. (If the property sets in M are ordered, M_- is the result of removing from M every property set s_n such that for some s_m in M , (a) $s_m < s_n$ and (b) s_m and s_n have the same exponence across all conjugations.) Accordingly, M'_- is the set $\{N: N \subseteq M_- \text{ and the exponence in } P_L \text{ of the morphosyntactic property sets belonging to } N \text{ suffices to determine the exponence in } P_L \text{ of every morphosyntactic property set belonging to } M_-\}$, and L 's paradigm predictability PP_L is calculated as in (10) rather than as in (8).

$$(10) \quad PP_L = \frac{|M'_-|}{|\mathcal{P}(M_-)|}$$

The second refinement in the calculation of paradigm predictability stems from the fact that where N is a large subset of M_- , the exponence in P_L of the morphosyntactic property sets belonging to N is generally very likely to determine the exponence in P_L of every morphosyntactic property set belonging to M_- . That is, the subsets of M_- that are best for distinguishing degrees of paradigm predictability tend to be the smaller subsets of M_- . We have therefore chosen—somewhat arbitrarily—to base our calculation of paradigm predictability on subsets of M_- having no more than seven members. For any set S of sets, we use $\leq_7 S$ to represent the largest subset of S such that for every $s \in \leq_7 S$, $|s| \leq 7$. We accordingly calculate L 's paradigm predictability PP_L as in (11) rather than as in (10).

$$(11) \quad \text{PPL} = \frac{|\leq_7 M'_-|}{|\leq_7 \mathcal{P}(M_-)|}$$

This measure of paradigm predictability accounts for both of the practical criteria in (6). Consider first criterion (6a), which associates a smaller number of dynamic principal parts with greater paradigmatic transparency. By this criterion, Conjugation P3A exhibits greater paradigmatic transparency than Conjugation P3J, since the only optimal analysis of Conjugation P3A involves a single principal part (Table 2.26), while the only optimal analysis for Conjugation P3J involves two principal parts (Table 2.27). This difference reflects a measurable contrast in the paradigm predictability of the two conjugations: the predictability of a member of Conjugation P3A is 0.450, while that of a member of P3J is merely 0.193.

Consider now criterion (6b), which associates greater paradigmatic transparency with a greater number of alternative inflection-class analyses. By this criterion, Conjugation PDBB exhibits greater paradigmatic transparency than Conjugation P3A, since the former allows the twelve optimal principal-part analyses in Table 2.11, while the latter only allows the single optimal principal-part analysis in Table 2.26. This difference reflects a measurable contrast in the paradigm predictability of these two conjugations: the predictability of a member of Conjugation PDBB is 1.000, while the predictability of a member of P3A is merely 0.450.

Applying the measure of paradigm predictability to all of the conjugations in Comaltepec Chinantec yields the results in Table 2.28 (represented

TABLE 2.26 The sole optimal principal-part analysis for Conjugation P3A in Comaltepec Chinantec

Principal part	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>	<i>11</i>

TABLE 2.27 The sole optimal principal-part analysis for Conjugation P3J in Comaltepec Chinantec

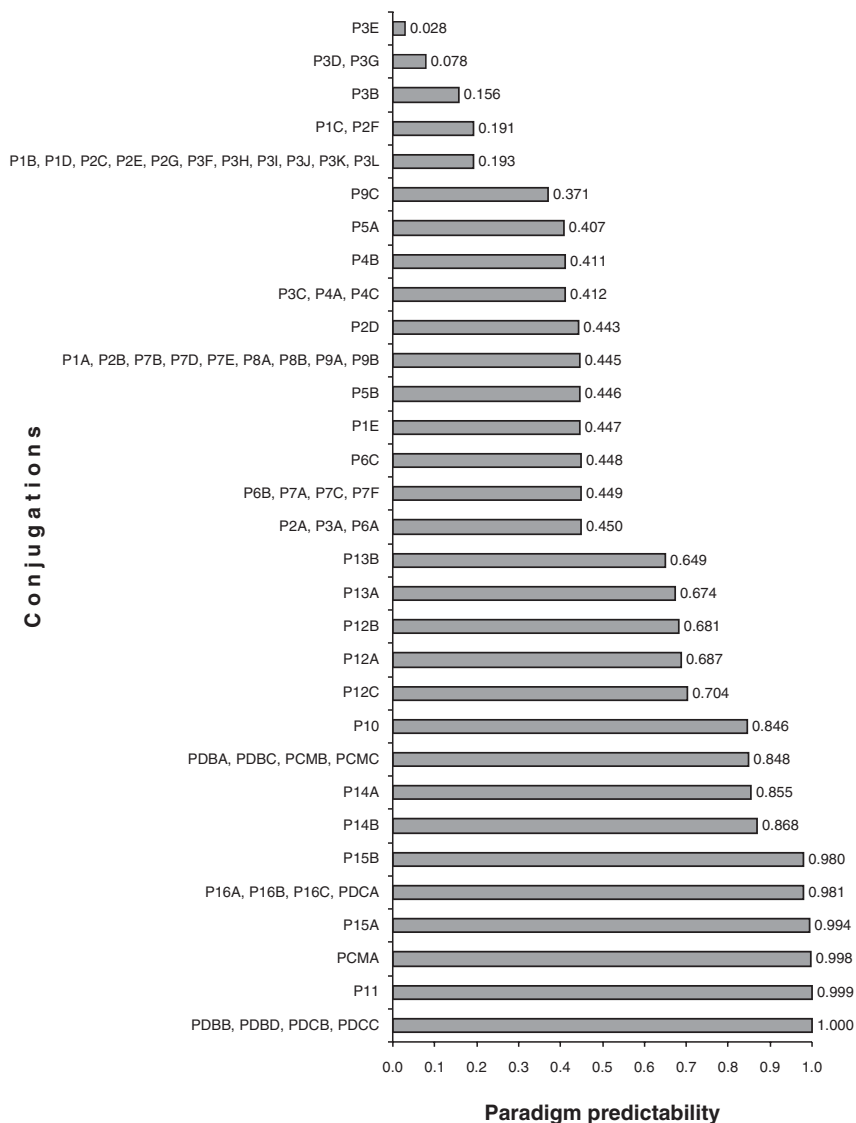
Principal part	Morphosyntactic property sets											
	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>
<i>9,11</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>9</i>	<i>11</i>	<i>9</i>

TABLE 2.28 Paradigm predictability across conjugations in Comaltepec Chinantec

Conjugations	Paradigm predictability
A. PDBB, PDBD, PDCB, PDCC	1.000
P ₁₁	0.999
PCMA	0.998
P _{15A}	0.994
P _{16A} , P _{16B} , P _{16C} , PDCA	0.981
P _{15B}	0.980
P _{14B}	0.868
P _{14A}	0.855
PDBA, PDBC, PCMB, PCMC	0.848
P ₁₀	0.846
P _{12C}	0.704
P _{12A}	0.687
P _{12B}	0.681
P _{13A}	0.674
P _{13B}	0.649
B. P _{2A} , P _{3A} , P _{6A}	0.450
P _{6B} , P _{7A} , P _{7C} , P _{7F}	0.449
P _{6C}	0.448
P _{1E}	0.447
P _{5B}	0.446
P _{1A} , P _{2B} , P _{7B} , P _{7D} , P _{7E} , P _{8A} , P _{8B} , P _{9A} , P _{9B}	0.445
P _{2D}	0.443
P _{3C} , P _{4A} , P _{4C}	0.412
P _{4B}	0.411
P _{5A}	0.407
P _{9C}	0.371
C. P _{1B} , P _{1D} , P _{2C} , P _{2E} , P _{2G} , P _{3E} , P _{3H} , P _{3I} , P _{3J} , P _{3K} , P _{3L}	0.193
P _{1C} , P _{2F}	0.191
P _{3B}	0.156
D. P _{3D} , P _{3G}	0.078
E. P _{3E}	0.028

graphically in Table 2.28a). Close inspection reveals four points at which the gradient of paradigm predictability in Table 2.28 breaks sharply; these breaks are the boundaries between parts A through E of the table. We believe that these breaks are best understood with respect to a second measure pertinent to paradigmatic transparency. CELL PREDICTABILITY measures the predictability of a cell's realization from the realization of the other cells in its paradigm (whether or not these are optimal principal parts).

Table 2.28a Paradigm predictability across conjugations in Comaltepec Chinantec



The fundamental idea underlying our proposed measure of cell predictability is that where (i) M is the set of morphosyntactic property sets associated with the cells in the paradigm P_L of some lexeme L and (ii) M_s is the set $\{N: N \subseteq M_- \text{ and the exponence in } P_L \text{ of the morphosyntactic property sets belonging to } N \text{ suffices to determine the exponence in } P_L \text{ of the property set } s\}$, the cell predictability $\text{CP}_{s,L}$ of s in P_L is calculated as in (12).

$$(12) \quad \text{CP}_{s,L} = \frac{|\leq_7 M_s|}{|\leq_7 \mathcal{P}(M_-)|}$$

Here, too, a refinement must be made. Because the exponence of property set s always suffices to determine itself, the inclusion of s in M_s invariably enhances cell predictability, thereby diminishing distinctions in cell predictability. We therefore exclude s from M_s in calculating cell predictability. For any collection C of sets, we use $C_{[s]}$ to represent the largest subset of C such that no member of $C_{[s]}$ contains s . Cell predictability is then calculated as in (13).

$$(13) \quad \text{CP}_{s,L} = \frac{|\leq_7 M_{s[s]}|}{|\leq_7 \mathcal{P}(M_-)_{[s]}|}$$

By this measure, the cells in the paradigms of Comaltepec Chinantec verbs have the cell predictability in Table 2.29; average cell predictability and paradigm predictability are listed in the table's rightmost two columns. The measure of cell predictability shows that the major breaks in the gradient of paradigm predictability correspond to the appearance of an unpredictable cell (i.e. one whose cell predictability is 0). The conjugations in part A of Table 2.28 have no unpredictable cells; those in part B have one unpredictable cell; those in part C have two unpredictable cells; and so on. (The cell predictability measures of unpredictable cells are shaded in Table 2.29.) Thus, the cell predictability measure reveals an important fact about paradigmatic transparency: Cell unpredictability degrades paradigm predictability. Inevitably, an unpredictable cell must be a principal part. (Table 2.29 is represented graphically in Fig. 2.2, in which morphosyntactic property sets are listed on the horizontal axis, conjugations are listed on the vertical axis (in order of decreasing paradigm predictability), and the lightness of a cell's shading represents its degree of cell predictability.)

2.6 Paradigmatic transparency and the No-Blur Principle

There can be no doubt that paradigmatic transparency helps the language user, both in the domain of language learning and in that of lexical storage.

TABLE 2.29 Cell predictability in all conjugations in Comaltepec Chinantec

Conj	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>11</u>	<u>12</u>	Avg cell predictability	Paradigm predictability
PDBB, PDBD, PDCB, PDCC	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.000
P11	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
PCMA	0.999	0.999	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.998	0.997	0.998	0.998
P15A	0.989	0.989	0.989	0.989	0.989	0.989	0.989	0.991	0.998	0.995	0.993	0.994	0.991	0.994
P16A, P16C	0.999	0.965	0.965	0.999	0.999	0.965	0.965	0.999	0.999	0.965	0.965	0.999	0.982	0.981
P16B, PDCA	0.999	0.999	0.965	0.965	0.999	0.999	0.965	0.965	0.999	0.999	0.965	0.965	0.982	0.981
P15B	0.964	0.964	0.964	0.964	0.997	0.997	0.964	0.965	0.999	0.997	0.989	0.990	0.979	0.980
P14B	0.892	0.892	0.892	0.892	0.974	0.974	0.974	0.991	0.859	0.857	0.798	0.798	0.900	0.868
P14A	0.990	0.990	0.990	0.990	0.860	0.860	0.860	0.998	0.858	0.737	0.807	0.859	0.900	0.855
PDBA, PDBC, PCMB	0.999	0.999	0.999	0.724	0.999	0.999	0.999	0.724	0.999	0.999	0.999	0.724	0.930	0.848
PCMC	0.997	0.997	0.997	0.724	0.998	0.997	0.997	0.724	0.997	0.998	0.860	0.724	0.917	0.848
P10	0.996	0.996	0.996	0.724	0.996	0.996	0.996	0.720	0.996	0.996	0.998	0.724	0.927	0.846
P12C	0.956	0.963	0.960	0.954	0.987	0.987	0.987	0.987	0.987	0.980	0.466	0.461	0.890	0.704
P12A	0.926	0.930	0.928	0.925	0.991	0.991	0.991	0.991	0.989	0.986	0.467	0.432	0.879	0.687
P12B	0.914	0.928	0.927	0.913	0.979	0.979	0.979	0.979	0.989	0.973	0.428	0.463	0.871	0.681
P13A	0.965	0.965	0.961	0.961	0.930	0.930	0.896	0.961	0.982	0.930	0.432	0.467	0.865	0.674
P13B	0.991	0.991	0.983	0.983	0.860	0.860	0.852	0.983	0.930	0.860	0.432	0.463	0.849	0.649
P2A	0.979	0.990	0.982	0.986	0.980	0.986	0.979	0.993	0.988	0.985	0.000	0.466	0.859	0.450
P3A	0.999	0.999	0.930	0.930	0.930	0.930	0.930	0.930	0.467	0.467	0.000	0.467	0.748	0.450
P6A	0.994	0.994	0.994	0.994	0.997	0.998	0.998	0.994	0.994	0.994	0.000	0.995	0.912	0.450
P6B	0.994	0.994	0.994	0.994	0.996	0.998	0.997	0.994	0.994	0.994	0.000	0.994	0.912	0.449
P7A	0.988	0.988	0.988	0.995	0.988	0.988	0.988	0.998	0.990	0.996	0.000	0.995	0.908	0.449
P7C	0.988	0.988	0.988	0.995	0.988	0.988	0.988	0.997	0.990	0.996	0.000	0.995	0.908	0.449
P7F	0.988	0.988	0.988	0.995	0.988	0.988	0.988	0.998	0.991	0.996	0.000	0.995	0.908	0.449
P6C	0.991	0.991	0.991	0.991	0.995	0.998	0.998	0.991	0.991	0.991	0.000	0.991	0.910	0.448
P1E	0.983	0.983	0.981	0.981	0.981	0.994	0.987	0.989	0.989	0.991	0.000	0.461	0.860	0.447
P5B	0.982	0.982	0.982	0.982	0.917	0.929	0.927	0.982	0.965	0.921	0.000	0.970	0.878	0.446
P1A	0.976	0.976	0.974	0.974	0.974	0.993	0.980	0.988	0.989	0.992	0.458	0.000	0.856	0.445

P2B	0.970	0.988	0.972	0.985	0.971	0.976	0.970	0.993	0.980	0.976	0.000	0.465	0.854	0.445
P7B	0.980	0.980	0.980	0.995	0.980	0.980	0.980	0.997	0.982	0.996	0.000	0.995	0.904	0.445
P7D	0.981	0.981	0.981	0.997	0.981	0.981	0.981	0.998	0.982	0.998	0.000	0.997	0.905	0.445
P7E	0.980	0.980	0.980	0.994	0.980	0.980	0.980	0.997	0.982	0.996	0.000	0.994	0.904	0.445
P8A	0.982	0.998	0.981	0.996	0.981	0.981	0.981	0.998	0.981	0.981	0.000	0.993	0.905	0.445
P8B	0.982	0.998	0.980	0.994	0.980	0.980	0.980	0.996	0.980	0.980	0.000	0.992	0.904	0.445
P9A	0.980	0.980	0.980	0.856	0.982	0.980	0.980	0.859	0.980	0.996	0.000	0.856	0.869	0.445
P9B	0.981	0.981	0.981	0.858	0.982	0.981	0.981	0.860	0.981	0.998	0.000	0.858	0.870	0.445
P2D	0.967	0.988	0.972	0.981	0.968	0.974	0.967	0.989	0.980	0.973	0.000	0.461	0.852	0.443
P3C	0.998	0.998	0.860	0.860	0.929	0.929	0.929	0.860	0.466	0.466	0.000	0.397	0.724	0.412
P4A	0.996	0.996	0.857	0.858	0.988	0.988	0.988	0.858	0.928	0.958	0.000	0.858	0.856	0.412
P4C	0.996	0.996	0.857	0.858	0.993	0.993	0.993	0.858	0.928	0.962	0.000	0.858	0.858	0.412
P4B	0.994	0.994	0.855	0.858	0.991	0.991	0.991	0.856	0.961	0.960	0.000	0.858	0.859	0.411
P5A	0.980	0.980	0.980	0.980	0.844	0.860	0.858	0.980	0.963	0.856	0.000	0.972	0.854	0.407
P9C	0.980	0.980	0.980	0.721	0.982	0.980	0.980	0.724	0.980	0.996	0.000	0.721	0.835	0.371
P1B	0.985	0.985	0.983	0.983	0.983	0.996	0.989	0.989	0.989	0.994	0.000	0.000	0.823	0.193
P1D	0.989	0.989	0.987	0.987	0.987	0.996	0.990	0.993	0.993	0.994	0.000	0.000	0.826	0.193
P2C	0.972	0.986	0.979	0.980	0.975	0.980	0.972	0.988	0.986	0.979	0.000	0.000	0.816	0.193
P2E	0.976	0.987	0.979	0.983	0.979	0.983	0.976	0.992	0.986	0.982	0.000	0.000	0.819	0.193
P2G	0.969	0.978	0.975	0.971	0.971	0.982	0.969	0.988	0.991	0.981	0.000	0.000	0.814	0.193
P3F	0.998	0.998	0.858	0.858	0.930	0.930	0.930	0.858	0.467	0.467	0.000	0.000	0.691	0.193
P3H	0.999	0.999	0.930	0.930	0.930	0.930	0.930	0.930	0.000	0.688	0.000	0.688	0.746	0.193
P3I	0.999	0.999	0.930	0.930	0.965	0.965	0.965	0.930	0.000	0.724	0.000	0.467	0.739	0.193
P3J	0.999	0.999	0.964	0.964	0.964	0.964	0.964	0.964	0.000	0.467	0.000	0.467	0.726	0.193
P3K	0.998	0.998	0.964	0.964	0.929	0.929	0.929	0.964	0.467	0.000	0.467	0.000	0.718	0.193
P3L	0.997	0.997	0.963	0.963	0.929	0.929	0.929	0.963	0.724	0.000	0.467	0.000	0.738	0.193
P1C	0.976	0.976	0.974	0.974	0.974	0.994	0.988	0.980	0.980	0.993	0.000	0.000	0.817	0.191
P2F	0.960	0.977	0.974	0.963	0.963	0.974	0.960	0.980	0.990	0.973	0.000	0.000	0.810	0.191
P3B	0.999	0.999	0.860	0.860	0.860	0.860	0.860	0.860	0.000	0.397	0.000	0.397	0.662	0.156
P3D	0.999	0.999	0.860	0.860	0.930	0.930	0.930	0.860	0.000	0.467	0.000	0.000	0.653	0.078
P3G	0.997	0.997	0.928	0.928	0.929	0.929	0.929	0.928	0.467	0.000	0.000	0.000	0.669	0.078
P3E	0.998	0.998	0.929	0.929	0.929	0.929	0.929	0.929	0.000	0.000	0.000	0.000	0.631	0.028

Nevertheless, the facts presented above raise doubts about the extent to which paradigmatic transparency is necessary in human language. In particular, they cast doubt on the No-Blur Principle, a hypothesis which portrays the avoidance of paradigmatic opacity as a structural principle of natural language.

Cameron-Faulkner and Carstairs-McCarthy (2000: 816) formulate the No-Blur Principle as in (14).

(14) The No-Blur Principle

Among the rival affixes for any inflectional cell, at most one affix may fail to be a class-identifier, in which case that one affix is the class-default for that cell.

This principle entails that all of the affixal exponents for the inflection of lexemes belonging to a particular category fall into two classes: class-identifiers and class-defaults.

- (15) a. A CLASS-IDENTIFYING affix is one that is peculiar to one inflection class, so that it can be taken as diagnostic of membership in that class.
 b. A CLASS-DEFAULT affix is one that is shared by more than one inflection class, and all of whose rivals (if any) are class-identifiers.

(Cameron-Faulkner and Carstairs-McCarthy 2000: 815)

If all affixes have to be either class-identifiers or class-defaults (as the No-Blur Principle assumes), then any lexeme that ever inflects by means of a class-identifier needs only one principal part: the word containing that class-identifier suffices to indicate which inflection class the lexeme belongs to. The only situation in which this won't hold true is one in which none of the words in a lexeme's paradigm contains a class-identifier; in that case, the lexeme's words must inflect entirely by means of class-default affixes. But if at most one affix per cell may fail to be a class-identifier, then there can only be one inflection class whose inflection is based entirely on class-default affixes. This, therefore, is the only inflection class whose members could have more than one principal part. That is, the No-Blur Principle has the entailment in (16):

- (16) Of all the inflection classes for lexemes of a given syntactic category, at most one requires more than one principal part.

The No-Blur Principle is apparently disconfirmed by Comaltepec Chinantec; but Cameron-Faulkner and Carstairs-McCarthy assume that the No-Blur Principle only relates to affixal exponence, and in Comaltepec Chinantec,

TABLE 2.30 Affixal and nonaffixal exponents of Fur conjugations

Conj	Examples ¹	Third person											
		Nonthird person			Singular			Plural					
		Subj	Perf	Pres	Subj	Perf	Pres	Nonhuman			Human		
Subj	Perf	Pres	Subj	Perf	Pres	Subj	Perf	Pres	Subj	Perf	Pres		
I,1a	buuN ‘descend’	LH-o	LH-ò	LH-èl	HH-o	HH-ò	HH-èl	HH-òl	HH-ùl	HH-èl-à/-i	LH-òl	LH-ùl	LH-èl-à/-i
I,1b	jaan ‘wait’	LH-o	LH-ò	LF-Ø	HH-o	HH-ò	HF-Ø	HH-òl	HH-ùl	HH-è	LH-òl	LH-ùl	LH-è
I,1c	irt ‘shake’	LH-o	LH-ò	LH-ì	HH-o	HH-ò	HH-ì	HH-òl	HH-ùl	HH-è	LH-òl	LH-ùl	LH-è
I,2a	tall ‘chew’	HH-ò	HH-o	HH-èl	LL-o	LL-ò	LL-èl	LL-òl	LL-ùl	LL-èl-à/-i	HH-òl	HH-ùl	HH-èl-à/-i
I,2b	fuul ‘spin’	HH-ò	HH-o	HF-Ø	LL-o	LL-ò	LL-Ø	LL-òl	LL-ùl	LL-è	HH-òl	HH-ùl	HH-è
I,2c	kir ‘cook’	HH-ò	HH-o	HH-ì	LL-o	LL-ò	LL-ì	LL-òl	LL-ùl	LL-è	HH-òl	HH-ùl	HH-è
II,1a	rii ‘snatch’	LH-i	LH-i	LH-itì	HH-i	HH-i	HH-itì	HH-i-A(1)	HH-i-è	HH-itì-A(1)	LH-i-A(1)	LH-i-è	LH-itì-A(1)
II,1b	tiir ‘meet’	LH-i	LH-i	LF-Ø	HH-i	HH-i	HF-Ø	HH-i-A(1)	HH-i-è	HH-è	LH-i-A(1)	LH-i-è	LH-è
II,2a	*faul ‘open’	HH-ì	HH-ì	HH-itì	LL-i	LL-i	LL-itì	LL-i-A(1)	LL-i-è	LL-itì-A(1)	HH-i-A(1)	HH-i-è	HH-itì-A(1)
II,2b	*kaun ‘grind’	HH-ì	HH-ì	HF-Ø	LL-i	LL-i	LF-Ø	LL-i-A(1)	LL-i-è	LL-è	HH-i-A(1)	HH-i-è	HH-è
IIIa	arr ‘measure’	HH-ì	HH-à	HH-èl	LH-ì	LH-à	LH-èl	LH-è	LH-e	LH-èl-à	HH-è	HH-e	HH-èl-à
IIIb	awi ‘pound’	HH-ò	HH-ò	HH-èl	LH-ò	LH-ò	LH-èl	LH-è	LH-e	LH-èl-à	HH-è	HH-e	HH-èl-à
IIIc	dus ‘tear’ (tr)	HH-ò	HH-ò	HH-èl	LF-Ø	LH-ò	LH-èl	LH-è	LH-e	LH-èl-à	HH-è	HH-e	HH-èl-à
IIId	*kair ‘stop’ (itr)	HF-Ø	HH-à	HH-èl	LF-Ø	LH-à	LH-èl	LH-è	LH-e	LH-èl-à	HH-è	HH-e	HH-èl-à
IIIe	*tai ‘hold, seize’	HF-Ø	HH-à	HH-èl	LF-Ø	LH-à	LH-èl	LH-è	LH-e	LH-èl-à	HH-è	HH-e	HH-èl-à
IVa	jum ‘cover’	HF-Ø	HH-ò	HH-èl	LF-Ø	LH-ò	LH-èl	LH-Al	LH-e	LH-èl-à	HH-Al	HH-e	HH-èl-à
IVb	bul ‘find’	HH-ò	HH-ò	HH-èl	LH-ò	LH-ò	LH-èl	LH-Al	LH-e	LH-èl-à	HH-Al	HH-e	HH-èl-à
IVc	juuN ‘terrify’	HF-Ø	HH-à	HH-èl	LF-Ø	LH-à	LH-èl	LH-Al	LH-e	LH-èl-à	HH-Al	HH-e	HH-èl-à
IVd	kur ‘touch’	HH-à	HH-à	HH-èl	LH-à	LH-à	LH-èl	LH-Al	LH-e	LH-èl-à	HH-Al	HH-e	HH-èl-à

Shaded cells represent dynamic principal parts in one optimal principal-part analysis.

1. The root forms in this column exclude tone markings.

(Source: Jakobi 1990: 103–13)

TABLE 2.31 Affixal exponents of Fur conjugations

Conj	Examples ¹	Third person											
		Nonthird person			Singular			Plural					
		Subj	Perf	Pres	Subj	Perf	Pres	Nonhuman			Human		
Subj	Perf	Pres	Subj	Perf	Pres	Subj	Perf	Pres	Subj	Perf	Pres		
I,1a	buuN 'descend'	-o	-ò	-èl	-o	-ò	-èl	-òl	-ùl	-èl-à/-i	-òl	-ùl	-èl-à/-i
I,1b	jaan 'wait'	-o	-ò	∅	-o	-ò	∅	-òl	-ùl	-è	-òl	-ùl	-è
I,1c	irt 'shake'	-o	-ò	-i	-o	-ò	-i	-òl	-ùl	-è	-òl	-ùl	-è
I,2a	tall 'chew'	-ò	-o	-èl	-o	-ò	-èl	-òl	-ùl	-èl-à/-i	-òl	-ùl	-èl-à/-i
I,2b	fuul 'spin'	-ò	-o	∅	-o	-ò	∅	-òl	-ùl	-è	-òl	-ùl	-è
I,2c	kir 'cook'	-ò	-o	-i	-o	-ò	-i	-òl	-ùl	-è	-òl	-ùl	-è
II,1a	rii 'snatch'	-i	-i	-iti	-i	-i	-iti	-i-A(l)	-i-è	-iti-A(l)	-i-A(l)	-i-è	-iti-A(l)
II,1b	tiir 'meet'	-i	-i	∅	-i	-i	∅	-i-A(l)	-i-è	-è	-i-A(l)	-i-è	-è
II,2a	*faul 'open'	-i	-i	-iti	-i	-i	-iti	-i-A(l)	-i-è	-iti-A(l)	-i-A(l)	-i-è	-iti-A(l)
II,2b	*kaun 'grind'	-i	-i	∅	-i	-i	∅	-i-A(l)	-i-è	-è	-i-A(l)	-i-è	-è
IIIa	arr 'measure'	-i	-à	-èl	-i	-à	-èl	-è	-e	-èl-à	-è	-e	-èl-à
IIIb	awi 'pound'	-ò	-ò	-èl	-ò	-ò	-èl	-è	-e	-èl-à	-è	-e	-èl-à
IIIc	dus 'tear' (tr)	-ò	-ò	-èl	∅	-ò	-èl	-è	-e	-èl-à	-è	-e	-èl-à
IIId	*kair 'stop' (itr)	∅	-à	-èl	∅	-à	-èl	-è	-e	-èl-à	-è	-e	-èl-à
IIIe	*tai 'hold, seize'	∅	-à	-èl	∅	-ò	-èl	-è	-e	-èl-à	-è	-e	-èl-à
IVa	jum 'cover'	-∅	-ò	-èl	∅	-ò	-èl	-Al	-e	-èl-à	-Al	-e	-èl-à
IVb	bul 'find'	-ò	-ò	-èl	-ò	-ò	-èl	-Al	-e	-èl-à	-Al	-e	-èl-à
IVc	juuN 'terrify'	∅	-à	-èl	∅	-à	-èl	-Al	-e	-èl-à	-Al	-e	-èl-à
IVd	kur 'touch'	-à	-à	-èl	-à	-à	-èl	-Al	-e	-èl-à	-Al	-e	-èl-à

Only the two affixal exponents in heavy boxes are class-identifiers.

Shaded cells represent dynamic principal parts in one optimal principal-part analysis.

1. The root forms in this column exclude tone markings.

(Source: Jakobi 1990: 103–13)

TABLE 2.32 Number of dynamic principal parts needed to identify each Fur conjugation

Conjugation	Number of dynamic principal parts	
	With only affixes taken into account	With tonality and affixes both taken into account
IIIa; IVd	1 (class-identifier)	1
I,1a; I,1c; I,2a; I,2b; I,2c; II,1a; II,2a; II,2b	2	1
I,1b; II,1b; IIIb; IIIc; IIIe; IVa; IVb	2	2
IIIc; IVc	3	3

conjugation classes are distinguished by non-affixal morphology. What about affixal exponence?

The affixal inflection of Fur (Nilo-Saharan; Sudan) decisively disconfirms the No-Blur Principle. In Fur, different conjugations are distinguished by the tonality of the verb root and by suffixation, as in Table 2.30.

Whether one takes account of the tonality of the root (as in Table 2.30) or not—that is, even if one restricts one’s attention purely to the affixes used in conjugation (as in Table 2.31)—there are nineteen conjugations in Fur.

The number of dynamic principal parts for a Fur conjugation class depends on whether one takes account of tonality. The two possibilities are given in Table 2.32. In this table, the lefthand column of numbers indicates the number of dynamic principal parts needed to identify each conjugation if only affixes are taken into account; the righthand column indicates the number required if root tonality as well as affixes are taken into account.

As the first column of Table 2.32 shows, only two of the nineteen conjugations have a class-identifier among their affixal exponents. By the assumptions of the No-Blur Principle, all of the other affixes in each column of Table 2.31 should be the class-default for that column; but this means that every one of the columns (= every morphosyntactic property set) in Table 2.31 has more than one class-default—contrary to the assumptions of the No-Blur Principle.

Cameron-Faulkner and Carstairs-McCarthy (2000) discuss an apparently similar instance from Polish in which a particular morphosyntactic property set (locative singular) seemingly has more than one class-default, namely the suffixes *-e* and *-u*. They argue, however, that these two suffixes actually constitute a single default, since they are in complementary distribution: *-e* only appears in combination with a lexeme’s special “minority” stem alternant,

and *-u* appears elsewhere. In this way, they claim, the Polish evidence can be reconciled with the No-Blur Principle.

This same strategy won't work for Fur, however. Notice, for example, that in the nonthird-person perfect, some conjugations exhibit a low-toned *-à* suffix and others exhibit a low-toned *-ò* suffix. Yet, the paradigms of conjugations exhibiting the *-à* suffix may exhibit exactly the same pattern of stem tonality as those of conjugations exhibiting the *-ò* suffix. For instance, Conjugations IIIe and IVa differ in that the first shows the *-à* suffix and the second shows the *-ò* suffix; yet, these two conjugations exhibit precisely the same pattern of stem tonality, and the two suffixes are therefore in contrastive rather than complementary distribution. More generally, for each of the six sets of conjugations listed in (17), the only differences in exponence between the conjugations are affixal, and none of the distinguishing affixes is a class-identifier. These facts lead inevitably to the conclusion that the No-Blur Principle cannot be maintained.

- (17) a. I-1a, I-1c and II-1a
 b. I-1b and II-1b
 c. I-2a, I-2c and II-2a
 d. I-2b and II-2b
 e. IIIb and IVb
 f. IIIc, IIIe, IVa and IVc

The theoretical antecedent of the No-Blur Principle is the Paradigm Economy Principle (Carstairs 1987), which Carstairs-McCarthy (1991: 222) formulates as in (18):

- (18) Paradigm Economy Principle
 There can be no more inflectional paradigms for any word-class in any language than there are distinct "rival" inflectional realizations available for that morphosyntactic property-combination where the largest number of rivals compete.

As with the No-Blur Principle, it is intended that this principle be interpreted as relating specifically to affixal inflection; thus, it entails that the maximum number of conjugations in Fur should be no larger than the maximum number of affixes that compete to realize the same property set in Fur verbal inflection. Just as the Fur evidence fails to confirm the predictions of the No-Blur Principle, it likewise fails to confirm the predictions of principle (18): in Fur, the largest number of "rival" suffixes for the inflection of a particular morphosyntactic property set is six (in both the nonthird-person subjunctive and the third-person singular subjunctive; cf. Table 2.31)—far fewer than the

total number of conjugations (of which there are nineteen). While the benefits of paradigm economy for language learning cannot be doubted, these facts show that paradigm economy is not clearly enforced by any grammatical constraint.

Accordingly, evidence from languages such as Fur and Comaltepec Chinantec raises similar doubts about Albright's (2002a: 11) single surface base hypothesis:

[T]he single base hypothesis means that for one form in the paradigm (the base), there are no rules that can be used to synthesize it, and memorization is the only option. Other forms in the paradigm may be memorized or may be synthesized, but synthesis must be done via operations on the base form. Since we are assuming here a word-based model of morphology, the base is a fully formed surface member of the paradigm, and for this reason, I will call this the *single surface base* hypothesis.

Albright acknowledges that in order to synthesize forms in a complex inflectional paradigm, it is sometimes necessary to refer to multiple, local bases; this might be taken to suggest that the paradigms of a richly inflected language can be subdivided into sectors such that each sector S has a base

TABLE 2.33 Degrees of transparency exhibited by Fur conjugations (with tonality as well as affixes taken into account)

Conjugation	Number of dynamic principal parts	Average number of principal parts needed to deduce a particular cell in a lexeme's paradigm	Number of optimal analyses	Paradigm predictability
I,1a ; II,1a ; II,2a	1	1.00	4	0.923
I,2a	1	1.00	3	0.922
II,2b	1	1.00	1	0.921
II,1b	2	1.00	32	0.918
I,1c ; I,2c ; IVd	1	1.00	2	0.707
IIIa	1	1.00	1	0.707
I,2b	1	1.00	1	0.706
I,1b	2	1.00	16	0.703
IVb	2	1.00	4	0.491
IVa	2	1.17	1	0.399
IVc	3	1.00	8	0.333
IIIb	2	1.00	2	0.309
IIIc ; IIIe	2	1.33	1	0.273
IIIId	3	1.00	4	0.206

by which the single surface base hypothesis is satisfied within S. But it's not clear that the single surface base hypothesis can be maintained even in this weakened form, since as we have seen, some of the forms in a paradigm are only deducible by simultaneous reference to two or more implicative forms within that paradigm. (See Finkel and Stump 2007 for additional relevant discussion.)

2.7 Paradigmatic transparency as a dimension of typological variation

Like the Comaltepec Chinantec facts, the Fur facts demonstrate that languages tolerate considerable variation in the amount of paradigmatic transparency that they exhibit. The relevant Fur facts are summarized in Table 2.33, where conjugations are distinguished according to four criteria: according to the number of dynamic principal parts required to characterize them, according to the average number of principal parts needed to deduce an individual cell in a lexeme's paradigm, according to the number of alternative optimal principal-part analyses available to them, and according to their paradigm predictability.

The measure of paradigm predictability reveals some significant typological contrasts between Comaltepec Chinantec and Fur. By this measure, Comaltepec Chinantec tolerates a lower degree of paradigmatic transparency than Fur does: more than a fourth of the conjugations in Comaltepec Chinantec have a paradigm predictability below 0.2, while none of the Fur conjugations has a paradigm predictability this low. This difference in tolerance is reflected in a number of ways. First, Comaltepec Chinantec has optimal analyses involving as many as four principal parts, in comparison with a maximum of three in Fur. Second, seventeen of the sixty-seven conjugations in Comaltepec Chinantec involve paradigms at least some of whose words have to be deduced by simultaneous reference to more than one principal part; in Fur, by contrast, only three of the nineteen conjugations involve paradigms some of whose words have to be deduced through simultaneous reference to more than one principal part. Third, Comaltepec Chinantec provides an example of a conjugation (namely P₃E) requiring four principal parts but allowing only one analysis out of a logically possible 495; Fur presents no conjugation class with a comparably constrained number of analyses. And fourth, well over half of the conjugations in Comaltepec Chinantec include one or more cells having a cell predictability of 0; by contrast, only four of the nineteen conjugations in Fur (namely IIIb, IIIc, IIId, and IIIe) have unpredictable

cells, and none has more than one unpredictable cell. Notwithstanding the fact that Comaltepec Chinantec clearly tolerates a lower degree of paradigmatic transparency than Fur, it does, at the same time, achieve maximal transparency in four conjugations, which no conjugation does in Fur.

2.8 Conclusions and projections for future research

Much past research on morphological typology has tended to focus on the structure of individual word forms, invoking such criteria as the average number of morphemes per word form and the degree of morpheme fusion within a word form. The criteria proposed here extend the focus of typological classification from the structure of individual word forms to that of whole paradigms and to the implicative relations that paradigms embody.

The principal-part analysis undertaken here dovetails with current probability-based research on the structure of inflectional paradigms (e.g. that of Ackerman, Blevins, and Malouf and Milin *et al.* in this volume). The latter work focuses on the probability that a given cell C in the paradigm of a given lexeme L of category G will have a given realization, where the factors affecting this probability include the number and relative frequency of the inflection classes to which members of G belong, the number and frequency of exponents competing for the realization of C across members of G, the realization of other cells in L's paradigm, and so on. The central measure of this probability is the information-theoretic notion of entropy: the higher a cell's ENTROPY, the less predictable (the more informative) its realization.

The notion of CONDITIONAL ENTROPY discussed by Ackerman, Blevins, and Malouf is particularly relevant to the notion of principal parts. If we already know the realization of cell A in some paradigm, that information may serve to diminish the entropy of cell B (i.e. to make its realization more predictable); this diminished entropy is the conditional entropy of B with respect to A. Where cell B belongs to a paradigm having A_1, \dots, A_n as its principal parts, A_1, \dots, A_n effectively reduce the entropy of cell B to zero (i.e. they make it fully predictable).

This, then, is the point of contact between principal-part analysis and probability-based research on paradigmatic structure: the former focuses on the number and identity of conditions that must be present in a paradigm in order to reduce the entropy of each of its cells to zero; in other words, it focuses not on the probability that a given cell has a given realization, but on the circumstances in which a cell's realization becomes a certainty. Thus, while the probability-based research of Ackerman, Blevins, and Malouf and Milin *et al.* is concerned with varying degrees of entropy in a language's paradigms,

our principal-part analyses are concerned with its varying degrees of paradigmatic transparency, i.e. the varying degrees of ease with which cells' realizations can be deduced with certainty from those of other cells in the same paradigm.

As we have shown, languages differ considerably in the extent to which they exhibit paradigmatic transparency. In view of the *prima facie* benefits of paradigmatic transparency for language learning and lexical storage, it is initially somewhat unexpected that languages should differ in this way. But paradigmatic transparency is by no means the only property of inflectional systems that may confer benefits on the language user. Transparadigmatic transparency—the ease with which a cell in one paradigm can be deduced from the corresponding cell in another paradigm – surely confers benefits of this sort; for instance, knowing that 1pl present indicative forms are alike across all conjugations makes the 1pl present indicative form of a newly learned verbal lexeme immediately deducible from those of existing lexemes. Yet the grammatical patterns that constitute paradigmatic transparency may be essentially the opposite of those constituting transparadigmatic transparency: a language all of whose conjugations possess maximal paradigmatic transparency (cf. again Fig. 2.1) possesses minimal transparadigmatic transparency; by the same token, a language in which distinct conjugation classes participated in a high degree of transparadigmatic transparency would inevitably exhibit low paradigmatic transparency. Thus, to understand the cross-linguistic variability of paradigmatic transparency, it will ultimately be necessary to understand the ways in which this property interacts with, counterbalances, or compensates for other, different grammatical properties.

Parts and wholes: Implicative patterns in inflectional paradigms

*Farrell Ackerman, James P. Blevins,
and Robert Malouf*

The whole has value only through its parts, and the parts have value only by virtue of their place in the whole. (Saussure 1916: 128)

... we cannot but conclude that linguistic form may and should be studied as types of patterning, apart from the associated functions. (Sapir 1921: 60)

3.1 Introduction

This chapter addresses an issue in morphological theory – and, ultimately, morphological learning – that we feel has received far less attention than it deserves. We will refer to this issue as the **Paradigm Cell Filling Problem** (PCFP):

Paradigm Cell Filling Problem: What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?

The problem does not arise in an isolating language, in which each lexical item (or “lexeme”) is realized by a single form. English, for all intents and purposes, approaches an isolating ideal, so that the PCFP has not been prominent in analyses of English (or, for that matter, in the post-Bloomfieldian morphological models that have been developed mainly within the English-speaking world).¹

However, the PCFP arises in an acute form in languages with complex inflectional systems, especially those which contain large inflectional paradigms

¹ Though a concern with form and structure of paradigms has remained a central focus of other morphological traditions, as represented by Seiler (1965), Wurzel (1970), and Carstairs (1983).

and intricate inflection-class systems. For example, a typical Estonian noun paradigm contains 30-odd forms, which exhibit patterns of variation that place the noun within anywhere between a half-dozen and a dozen major declension classes (Viks 1992; Erelt *et al.* 1995; Blevins 2005). It is implausible to assume that a speaker of Estonian will have encountered each form of every noun, so that native command of the language must involve the ability to generalize beyond direct experience. Moreover, Estonian is far from an extreme case. A typical transitive verb in Georgian has upwards of 200 forms, whose inflectional patterns identify the verb as belonging to one of four major conjugation classes (Tschenkéli 1958). Even Georgian is relatively conservative in comparison with descriptions of verb paradigms in Archi, which, according to one estimate (Kibrik 1998: 467), may contain “more than one and a half million” members.

The basic challenge that a speaker faces in each of these cases is the same, irrespective of the size of the form inventory. Given prior exposure to at most a subset of forms, how does a speaker produce or interpret a novel form of an item? One superficially attractive intuition is that knowing *what* one wants to say suffices in general to determine *how* one says it. The idea that variation in form reflects differences in “grammatical meaning” is encapsulated in the post-Bloomfieldian “morpheme,” and underlies morphemic models from Harris (1942) and Hockett (1947) through Lieber (1992) and Halle and Marantz (1993). Yet, if one thing has been established about morphological systems in the half-century since Hockett (1954), it is that complex systems exhibit genuinely morphological variation, which is not conditioned by differences in grammatical meaning (or, for that matter, solely by phonological factors). Purely morphological variation (or what Aronoff 1994 terms “morphology by itself”) may seem enigmatic in the context of simple systems. But in larger and more complex systems, variation that identifies the class of an item contributes information of vital importance because it allows a speaker to predict other forms of the item.

In a language with inflection classes, a speaker must be able to identify the class of an item in order to solve the PCFP. That is, to produce or interpret a novel form of an item, it is not enough for the speaker to know just that the grammatical meaning “motion into” is expressed by the illative case. The speaker must also know how the illative is realized for the item in question. In an inflection-class language, the choice of stem choice or exponent is precisely what is not in general determinable from the semantic or grammatical properties of an item. Instead, a speaker must know, or be able to deduce, one of the diagnostic forms of an item. For example, no known grammatical properties explain why the Estonian noun LUKK ‘lock’ has the short illative singular form *lukku* alongside the long form *lukusse*, whereas KIRIK ‘church’

has just the long form *kirikusse*. However, this contrast follows immediately if one knows that the partitive singular of LUKK is *lukku* and that the partitive singular of KIRIK is *kirikut*. There is likewise no morphosyntactic motivation for the variation in the form of the illative singulars in the Saami paradigms in Table 3.3 below. It is a morphological fact that the illative singular *bihttái* is based on the strong stem of BIHTTÁ ‘piece’ whereas the illative singular *bastii* is based on the weak stem of BASTE ‘spoon.’ This contrast is again predictable from the grade of the nominative singular forms of each noun (or, indeed, from the grade of any other form, as shown in Section 3.1).

In short, morphological systems exhibit interdependencies of precisely the kind that facilitate the deduction of new forms, based on knowledge of other forms. In some cases, it may be possible to mediate these deductions through a level of analysis in which recurrent units of form are associated with discrete grammatical meanings. However, this type of analysis tends to be most applicable to simple or recently grammaticalized patterns, and most morphological systems are not organized in a way that facilitates the identification of “minimal meaningful units”. In many cases, the interdependencies that hold between word forms do not hold between subword units, so that further analysis disrupts the implicational structure. For example, the partitive singular *lukku* implies the homophonous short illative singular *lukku*, even though neither *lukk* nor *-u* can be associated with the grammatical meaning “partitive” or “illative” (Blevins 2005).

To develop this perspective, Section 3.2 outlines the word and paradigm assumptions that underlie our analysis, together with the basic information theoretic measures we use to test these assumptions. In Section 3.3, we apply these measures to portions of the morphological systems of Saami and Finnish and argue that – even in the absence of accurate frequency information – these measures bring out an implicational structure that offers a solution to the PCFP. We then show how the same measures apply to a description of Tundra Nenets nouns that supplies information about type frequency.² Taken together, these case studies suggest how information theory can be used to measure the implicational relations that underlie **symmetrical** approaches to word relatedness. By measuring the information that multiple surface forms provide about other forms, these approaches capture patterns of interdependency that cannot always be expressed in terms of an **asymmetrical** relation between surface forms and a single underlying or surface base.³ We illustrate a symmetrical approach by examining Tundra Nenets nominal

² The fieldwork on Tundra Nenets was supported by a Hans Rausing Endangered Language Major Documentation Project Grant 2003–6, in which the first author was a co-PI with Irina Nikolaeva and Tapani Salminen. This support is gratefully acknowledged.

³ See Albricht (2002a, this volume) for a single base approach that addresses language change.

declension classes for absolute paradigms, and offer some provisional results about paradigm organization in this language. Section 3.4 then closes with some general conclusions and speculates about their ramifications for theoretical approaches to morphological analysis.

3.2 Analytical assumptions

Processes of analogical pattern matching and pattern extension play a central role in traditional analyses of interdependencies within and across paradigms. In classical word and paradigm (WP) models, a morphological system is factored into two components: a set of exemplary paradigms that exhibit the inflectional patterns of a language, and sets of diagnostic principal parts for nonexemplary items. Matching diagnostic forms of an item against the corresponding cells in an exemplary paradigm provides an analogical base for the deduction of novel forms of the item. This process of matching and deduction tends to be expressed symbolically in terms of proportional analogies (discussed in more detail in Albright (this volume) and Milin *et al.* (this volume)). The same process is invoked in grammars of inflectionally complex languages, as illustrated by the “rules of analogy” in Viks (1992: 46), which identify those forms of an Estonian noun that are predictable from the genitive singular and from the genitive plural.

3.2.1 Morphological assumptions

Traditional WP models offer a general solution to the PCFP that exploits the implicational structure of inflectional systems. Strategies that use exemplary patterns to extend principal part inventories are strikingly effective, as Matthews (1991: 187) notes in connection with their pedagogical relevance. They are also remarkably economical. In general, a small set of principal parts is sufficient to identify the class of an item and predict other forms of the item. Yet traditional solutions to the PCFP also raise some basic questions, including those in (1):

- (1) a. What is the structure of units that license implicative relations?
- b. How are units organized into larger structures within a system?
- c. How can one measure implicative relations between these units?
- d. How might the implicative organization of a system contribute to licensing inferences that solve the paradigm cell filling problem?
- e. How does this organization, and the surface inferences it licenses, contribute to the robustness and learnability of complex systems?

Questions (1a) and (1b) centrally concern the relation of parts to wholes along two independent dimensions of analysis. Question (1a) concerns the internal complexity of word forms. Within post-Bloomfieldian models, words are treated as aggregates of smaller meaningful elements. These parts combine to produce a whole whose meaning is just the sum of the meaning its parts. Within the WP approach adopted here, words are regarded as complex configurations of recurrent elements whose specific **patterns of combination** may be meaningful irrespective of whether any particular piece bears a discrete meaning.

From this perspective, a surface word form is a whole in which the patterns exhibited by parts – whether affixes, tones, ablaut, or other “features of arrangement” (Bloomfield 1933: 163) – merely signal morphosyntactic, lexical, or morphological properties.⁴ For example, in Tundra Nenets, the same members of a suffix set can be used with different lexical categories, sometimes serving essentially the same function, and sometimes serving different functions.⁵

As shown in Table 3.1, markers from Suffix Set I can appear both on nouns and verbs, and the inflected word functions as the predicate of the clause. In either case, the set I markers reflect person and number properties of the clausal subject. While markers from Suffix Set II also occur either with nouns or with verbs, their function differs within each class: they reflect person/number properties of the possessor when they appear with nouns, but number properties of clausal objects when they appear with (transitive) verbs. Hence, there is a configurational dynamic whereby the same elements in different combinations are associated with different meanings. These patterns show why words are best construed as **recombinant gestalts**, rather

⁴ This perspective does not preclude the possibility of associating grammatical meaning with subword units (morphemes) in constructions and/or languages where they would be motivated. In contrast, a morphemic model is less flexible, as it uniformly associates grammatical meaning with minimal elements and ignores configurational (emergent) properties of patterns.

⁵ This discussion follows the presentation in Salminen (1997: 96, 103, 126), though elsewhere we have simplified his transcriptions for a general audience. In section 3.3.3 we have largely rendered the traditional Cyrillic written conventions into an IPA-based system where digraphs such as *ny* indicate palatalized consonants, *ᵛ* refers to a glottal stop with nasalizing or voicing effects in sandhi contexts, and *ʔ* refers to a glottal stop without nasalizing effects in sandhi contexts. (For a detailed discussion of motivations for the specific orthographic symbols employed in exemplary word forms see Salminen (1993).) Also, while predicate nominals and adjectives in Tundra Nenets host markers from Suffix Set I, they differ from the verbal predicates that host these suffixes in exhibiting nominal stem formation rather than verbal stem formation, in the inability to host future markers, and in their manner of clausal negation. All of these differences suggest that two different lexical categories host markers from Suffix Set I, and that there is no N-to-V conversion operation.

TABLE 3.1 Suffix homonymy in Tundra Nenets

	N	V
Suffix Set I	Predicative	Subjective
Suffix Set II	Possessive	Objective

than simple (or even complex) combinations of bi-unique content-form mappings (i.e., morphemes).⁶

This perspective on complex words is intimated in Saussure (1916: 128) in his discussion of **associative** (= paradigmatic) relations:

A unit like **painful** decomposes into two subunits (**pain-ful**), both these units are not two independent parts that are simply lumped together (**pain + ful**), The unit is a product, a combination of two interdependent elements that acquire value only through their reciprocal action in a higher unit (**pain × ful**). The suffix is non-existent when considered independently; what gives it a place in the language is series of common terms like **delight-ful**, **fright-ful**, etc. . . . The whole has value only through its parts, and the parts have value by virtue of their place in the whole.

Accordingly, while we are often able to isolate pieces of complex form, it is the configurations in which these pieces occur and the relation of these configuration to other similar configurations that are the loci of the meanings that are relevant in morphology. This property becomes even more evident if one considers the structure of Tundra Nenets verbs as insightfully discussed and schematized in Salminen (1997).

Table 3.2 exhibits little in the way of a one-to-one correspondence between cells across columns. Consider first the general finite stem, whose use is exemplified in (2). This stem serves as the base for the subjective conjugation,

TABLE 3.2 More suffix homonymy in Tundra Nenets (Salminen 1997:96)

Conjugation	Number of Object	Morphological Substem	Suffix Set
subjective		general finite stem (modal substem)	I
	sg		II
objective	du	dual object (modal) substem	III
	pl		
reflexive		special modal stem	IV

⁶ See Gurevich (2006) for an constructional analysis of Georgian along these lines.

as shown in (2a), the objective conjugation, as shown in (2b), and may also encode singular object agreement for verbs marked by Suffix Set II. The dual object (modal) substem hosts members of Suffix Set III, as exemplified in (3), but the same suffix set also serves to mark plural objects with the special finite stem in (4a). Finally, as (4b) shows, the special finite stem is not restricted to the plural object conjugation, given that it is also associated with the reflexive conjugation and the distinguishing characteristic of this conjugation is the use of suffix set IV.

- (2) General finite stem:
- a. Subjective:
 tontaød⁰m
 cover.I (= 1sg)
 ‘I cover (something)’
 - b. Objective Singular:
 tontaøw⁰
 cover.II (= 1sg/sg)
 ‘I cover it’
- (3) Dual Object Stem:
 tontangax⁰ yun⁰
 cover.dual.III (= 1sg/du)
 ‘I cover them (two)’
- (4) Special finite stem:
- a. Objective Plural
 tonteyøn⁰
 cover.III (= 1sg/pl)
 ‘I cover them (plural)’
 - b. Reflexive
 tonteyøw⁰q
 cover.IV (= 1sg)
 ‘I got covered’

In sum, it is the pattern of arrangements of individual elements that realize the relevant lexical and morphosyntactic content associated with words that is important in these examples, rather than the sum of uniquely meaningful pieces.

A word-based perspective on these aspects of the **internal** organization of lexical units is highly compatible with a traditional conception of the second part-whole dimension, namely the **external** organization of words. In what Matthews (1991) below terms the “ancient model”, individual words function

as minimal elements in networks of elements, including inflectional paradigms, and paradigms are organized into larger networks, which include inflection classes.

In the ancient model the primary insight is not that words can be split into formatives, but they can be located in paradigms. They are not wholes composed of simple parts, but are themselves the parts within a complex whole. (Matthews 1991: 204)

The notions of internal and external structure are not exclusive – as they are sometimes thought to be – but, instead, represent complementary perspectives on a morphological system. Indeed, these two dimensions give rise to a paradigmatic variant of “duality of patterning” (Hockett 1960), in that they show how combinations of individually meaningless elements, whether morphs or other “features of arrangement”, compose words whose meaning depends in part on the place they occupy within larger paradigmatic structures. These complementary notions also permit an exploration of the intuitions evident in the twin themes of the epigrams above. In order to address these issues, the following sections explore how several Uralic languages (Saami, Finnish, and particularly Tundra Nenets) provide fertile ground for identifying the nature of the challenges posed by the PCFP, as well as the type of analysis best suited to address them.

Traditional WP approaches suggest answers to the other questions in (1), though in addressing these questions, it is important to separate the substantive claims and hypotheses of a WP model from any idealizations or simplifying assumptions introduced in the use of these models in reference or pedagogical grammars. For practical purposes, it is usually convenient in written grammars to represent lexical items by a single principal part wherever possible. Yet there is no reason to attribute any linguistic or psychological relevance to this extreme level of lexical economy. There are many well-described systems in which class can only be identified on the basis of multiple principle parts. Estonian conjugations provide a fairly straightforward illustration (Blevins 2007) as do the systems described in Finkel and Stump (this volume).⁷ From a psycholinguistic perspective, there is considerable evidence that frequency is, in fact, the primary determinant of whether a given form is stored in the mental lexicon of a speaker (Stemberger and MacWhinney 1986; Baayen *et al.* 2003*b*). Similarly, grammars tend to take the smallest diagnostic forms of an item as principal parts, even though any form

⁷ It may be significant that models incorporating something like the “single base hypothesis” (Albright 2002*a*, this volume) tend to be developed on the basis of comparatively simple systems.

(or set of forms) that identifies class is equally useful, and the choice of memorized forms is again likely to reflect frequency or other distributional properties rather than morphosyntactic or morphotactic properties.

Other issues that are implicit in traditional analogical models have been addressed in recent work. Methods for identifying and classifying principal part inventories are set out in Finkel and Stump (2007, this volume). The psychological status of proportional analogies is likewise addressed in Milin *et al.* (this volume). But traditional solutions to the PCFP remain fundamentally incomplete to the extent that they lack a means of gauging the diagnostic value of principal parts or of measuring the implicational structure of networks of forms.

The approach outlined in this paper proceeds from the observation that implicational structure involves a type of **information**, specifically information that forms within a set convey about other forms in that set. Information in this sense corresponds to reduction in **uncertainty**. The more informative a given form is about a set of forms, the less uncertainty there is about the other forms in the set. The PCFP just reflects the fact that a speaker who has not encountered all of the forms of a given item is faced with some amount of uncertainty in determining the unencountered forms. If the choice of each form were completely independent, the PCFP would reduce to the problem of learning the lexicon of an isolating language. However, in nearly all inflectional systems, there are at least some forms of an item that reduce uncertainty about the other forms of the item. It is the reduction in uncertainty due to the knowledge of these forms that defines the implicational structure of the system. The diagnostic value of a given form likewise correlates with the reduction in uncertainty that is attributable to the knowledge of this particular form. Once these notions are construed in terms of uncertainty reduction, the task of measuring implicational structure and diagnostic value is susceptible to well-established techniques of analysis.

3.2.2 *Information theoretic assumptions*

The uncertainty associated with the realization of a paradigm cell correlates with its **entropy** (Shannon 1948) and the entropy of a paradigm is the sum of the entropies of its cells. The implicational relation between a paradigm cell and a set of cells is modeled by **conditional entropy**, the amount of uncertainty about the realization of the set that remains once the realization of the cell is known. Finally, the diagnostic value of a paradigm cell correlates with the **expected conditional entropy** of the cell, the average uncertainty that remains in the other cells once the realization of the cell is known.

A straightforward application of these information-theoretic notions provides a natural means of measuring the implicational structure of inflectional systems. In particular, we use the notion of **information entropy** to quantify the uncertainty in the realization of a particular cell of a paradigm. As in Moscoso del Prado Martín *et al.* (2004), Milin *et al.* (2009) and Milin *et al.* (this volume), an information-theoretic perspective permits us to reconsider basic linguistic questions, in this case questions about the synchronic structure of inflectional systems.

In order to quantify the interrelations between forms in a paradigm, we use the information theoretic notion **entropy** as the measure of predictability. This permits us to quantify “prediction” as a change in uncertainty, or information entropy (Shannon 1948). The idea behind information entropy is deceptively simple: Suppose we are given a random variable X which can take on one of a set of alternative values x_1, x_2, \dots, x_n with probability $P(x_1), P(x_2), \dots, P(x_n)$. Then, the amount of uncertainty in X , or, alternatively, the degree of surprise we experience on learning the true value of X , is given by the entropy $H(X)$:

$$H(X) = - \sum_{x \in X} P(x) \log_2 P(x)$$

The entropy $H(X)$ is the weighted average of the **surprisal** $-\log_2 P(x_i)$ for each possible outcome x_i . The surprisal is a measure of the amount of information expressed by a particular outcome, measured in bits, where 1 bit is the information in a choice between two equally probable outcomes. Outcomes which are less probable (and therefore less predictable) have higher surprisal. Surprisal is 0 bits for outcomes which always occur ($P(x) = 1$) and approaches ∞ for very unlikely events (as $P(x)$ approaches 0). The more choices there are in a given domain and the more evenly distributed the probability of each particular occurrence, the greater the uncertainty or surprise there is (on average) that a particular choice will be made among competitors and, hence, the greater the entropy. Conversely, choices with only a few possible outcomes or with one or two highly probable outcomes and lots of rare exceptions have a low entropy.

For example, the entropy of a coin flip as resulting in either heads or tails is 1 bit; there is equal probability for an outcome of either heads or tails:

$$\begin{aligned} H(X) &= - \sum_{x \in X} P(x) \log_2 P(x) \\ &= -(P(h) \times \log_2 P(h) + P(t) \times \log_2 P(t)) \\ &= -(0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5) \\ &= 1 \end{aligned}$$

The entropy of a coin rigged to always come up heads, on the other hand, is 0 bits: there is no uncertainty in the outcome:

$$\begin{aligned}
 H(X) &= - \sum_{x \in X} P(x) \log_2 P(x) \\
 &= -(P(h) \times \log_2 P(h) + P(t) \times \log_2 P(t)) \\
 &= -(1.0 \times \log_2 1.0 + 0.0 \times \log_2 0.0) \\
 &= 0
 \end{aligned}$$

For other possible unfair coins, the entropy will fall somewhere between these extremes, with more biased coins having a lower entropy. We can extend this to find the **joint entropy** of more than one random variable. In general, the joint entropy of independent events is the sum of the entropies of the individual events. Suppose X is the outcome of one flip of a fair coin and Y is the outcome of a second flip. If the two flips are independent, then the probability of getting, say, heads on the first flip and tails on the second is the probability of getting heads on first times the probability of getting tails on the second, or $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. So, then, the joint entropy $H(X, Y)$ is:

$$\begin{aligned}
 H(X, Y) &= - \sum_{x \in X, y \in Y} P(x, y) \log_2 P(x, y) \\
 &= -(P(h, h) \times \log_2 P(h, h) + P(h, t) \times \log_2 P(h, t) \\
 &\quad + P(t, h) \times \log_2 P(t, h) + P(t, t) \times \log_2 P(t, t)) \\
 &= -4 \times (0.25 \times \log_2 0.25) \\
 &= 2
 \end{aligned}$$

3.3 Modeling implicational structure

With the previous section as background we can now measure the entropy of the inflectional systems mentioned earlier. In order to exhibit the general character of the PCFP and demonstrate how an information-theoretic approach calculates the relative diagnosticity of words, the following subsections present several morphological patterns with ascending levels of complexity. We first describe the basic patterns, restricting attention to instructive aspects of the organization of these systems, and then develop entropy-based analyses that reveal their implicational structure. The inflectional paradigms of Uralic languages are particularly instructive because of the way that they realize inflectional properties by distinctive combinations of stem alternations and affixal exponence. Hence these systems are not amenable to a standard

head-thorax-abdomen analysis in which lexical properties are expressed by the root, morphological class properties by stem formatives, and inflectional properties by inflectional affixes. For expositional convenience, we will initially assume, contrary to fact, that each cell in the paradigms below are equiprobable, so that speakers are just as likely to encounter one specific cell as any other.⁸ As will be shown in the following sections, an appealing property of an entropy-based measure of word relatedness is that they can be easily scaled up to data sets of increasing veridicality.

3.3.1 Northern Saami

Noun declensions in Northern Saami (Bartens 1989; Nickel 1990) offer a straightforward illustration of the PCFP. First-declension nouns, i.e., nouns whose stems have an even number of syllables, may inflect according to either of the patterns in Table 3.3. In nouns of the “weakening” type, the nominative and illative singular and the essive are all based on the strong stem of a noun, and the remaining forms are based on the weak stem. Nouns of the “strengthening” variety exhibit a mirror-image pattern, in which the nominative and illative singular and essive are based on the weak stem, and other forms are based on the strong stem. Strong forms, which are set in bold in Table 3.3, contain a geminate consonant which corresponds to a nongeminate in the corresponding weak forms.

On standard descriptions that recognize a single, number-neutral essive form, there are eleven cells in a first-declension paradigm. Hence, to solve the PCFP, a speaker must deduce at most ten forms. This task is greatly facilitated

TABLE 3.3 Gradation in first declension nouns in Saami (Bartens 1989: 511)

	‘Weakening’		‘Strengthening’	
	Sing	Plu	Sing	Plu
Nominative	bihttá	bihtát	baste	basttet
Gen/Acc	bihtá	bihtáid	bastte	basttiid
Illative	bihttái	bihtáide	bastii	basttiide
Locative	bihtás	bihtáin	basttes	basttiin
Comitative	bihtáin	bihtáiguin	basttiin	basttiiguin
Essive		bihttán		basten
		‘piece’		‘spoon’

⁸ Assuming equiprobable realizations also gives us an upper bound on the uncertainty in a paradigm. Since it is unlikely that all realizations are in fact equally likely, the actual entropy will almost always be lower than this.

TABLE 3.4 Invariant case endings in Saami
(*e* assimilates to *i* before *i*)

	Sing	Plu
Nominative	—	-t
Gen/Acc	—	-id
Illative	-i	-ide
Locative	-s	-in
Comitative	-in	-iguin
Essive		-n

by three general patterns. First, case endings are invariant, as illustrated in Table 3.4, so the endings can be memorized and need not be determined for individual first-declension nouns. Second, the comitative singular and locative plural are always identical, so a speaker must encounter at most one of these two forms. The third and most fundamental pattern relates to stem alternations. Given that endings are invariant, solving the PCFP for an item reduces to the problem of determining the distribution of strong and weak stems. This task is made much easier by the fact that the cells of a first-declension paradigm divide into the same two “cohort sets” in the weakening and strengthening patterns. Set A contains the nominative and illative singular and essive, and Set B contains the remaining cells. In nouns of the weakening type, Set A is strong and Set B is weak; in nouns of the strengthening type, Set A is weak and Set B is strong.

A striking consequence of this symmetry is that every form of a first-declension noun is diagnostic. A strong form from Set A identifies a noun as belonging to the weakening type, and licenses the deduction that the remaining Set A forms are strong and the Set B forms are weak. Conversely, a weak form from Set A identifies a noun as belonging to the strengthening type, and licenses the deduction that the remaining Set A forms are weak and the Set B forms are strong. Any Set B form, whether strong or weak, is equally diagnostic. In sum, knowing the form of any one paradigm cell eliminates nearly all uncertainty about the forms that fill the other cells in a first declension paradigm. This implicational structure is completely symmetrical. Each form of a paradigm is equally informative, and the nominative and accusative singular forms that realize noun stems play no privileged role in distinguishing noun types.

A straightforward application of information-theoretic notions provides a natural means of measuring the implicational structure of the Saami system. To measure the uncertainty of forms in an inflectional paradigm P , we let P be a matrix whose dimensions are defined by features, and a paradigm cell C be a variable which takes as values the different realizations of the features associated with C . If the entropy of each cell of the Saami paradigm is 1 bit, and there are eleven cells in the paradigm, then if all cells were independent we would expect the overall entropy of the paradigm (that is, the joint entropy of all the cells) to be 11 bits. However, there are only two subdeclensions in Table 3.3, and if we again assume that each is equally likely, then the overall entropy of the paradigm is also 1 bit. This shows that there is a significant amount of shared information in the Saami paradigm. In fact, once you know the realization of one cell, you know the realization of every other cell: any one cell completely predicts the others. One can quantify the degree of prediction between these cells using entropy. The average uncertainty in one variable given the value another is the **conditional entropy** $H(Y|X)$. If $P(y|x)$ is the conditional probability that $Y = y$ given that $X = x$, then the conditional entropy $H(Y|X)$ is:

$$H(Y|X) = - \sum_{x \in X} P(x) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$$

Conditional entropy can also be defined in terms of joint entropy:

$$H(Y|X) = H(X,Y) - H(X)$$

The smaller that $H(Y|X)$ is, the more predictable Y becomes on the basis of X , i.e., the less surprised one is that Y is selected. In the case where X completely determines Y , the conditional entropy $H(Y|X)$ is 0 bits: given the value of X , there is no question remaining as to what the value of Y is. On the other hand, if X gives us no information about Y at all, the conditional entropy $H(Y|X)$ is equal to $H(Y)$: given the value of X , we are just as uncertain about the value of Y as we would be without knowing X .

Given the paradigm in Table 3.3, we can calculate the conditional entropy of any one cell given any other cell. Let us take the nominative singular and the locative plural, which happen to belong to different cohort sets. Each cell has two possible realizations, and the entropy of each is 1 bit. To find the joint entropy, we look at the four possible combinations of realizations:

Nom Sg	Loc Pl	<i>P</i>
strong	strong	0.0
strong	weak	0.5
weak	strong	0.5
weak	weak	0.0

Once again, we have two equally likely possible outcomes, and the joint entropy is 1 bit. So, the conditional entropy is:

$$\begin{aligned}
 H(\text{LOC.PL}|\text{NOM.SG}) &= H(\text{NOM.SG}, \text{LOC.PL}) - H(\text{NOM.SG}) \\
 &= 1.0 - 1.0 \\
 &= 0.0
 \end{aligned}$$

That is, knowing the nominative singular realization for a particular noun completely determines the realization of the locative plural. One could repeat this calculation for any pair of cells in the paradigm and we would get the same result, as the Saami nominal inflection is a completely symmetric system.

In contrast, merely knowing one or both of the stem forms of a noun does not reduce uncertainty about whether a noun is of the weakening or strengthening type, because one must still know whether **which cell** the stem realizes. Knowing that the noun *BIHTTÁ* in Table 3.3 has the strong stem *bihttá* and the weak stem *bihtá* does not identify the subtype of this noun unless one knows which stem underlies which cohort set. Knowing that *BASTE* has the strong stem *bastte* and the weak stem *baste* is similarly uninformative. Hence, the type of these nouns cannot be determined from their stem inventories but only from the distribution of stems in the inflectional paradigms of the nouns.

3.3.2 Finnish

The Finnish subparadigm in Table 3.5 illustrates a more typical pattern, in which different **combinations** of cells are diagnostic of declension class membership.⁹ Although individual forms may be indeterminate with respect to class membership, particular combinations of forms in Table 3.5, varying from class to class, reduce the uncertainty of class assignment. Consider forms *laseissa*, *nalleissa* and *kirjeissa*, which realize the inessive plural in the paradigms of the nouns of *LASI*, *NALLE*, and *KIRJE*. None of these forms alone reliably predicts the corresponding nominative singular forms. But collectively

⁹ The numbers in Table 3.5 refer to the declension classes in Pihel and Pikamäe (1999).

TABLE 3.5 Finnish *i*-stem and *e*-stem nouns (Buchholz 2004)

Nom Sg	Gen Sg	Part Sg	Part Pl	Iness Pl	
ovi	oven	ovea	ovia	ovissa	'door' (8)
kieli	kielen	kieltä	kieliä	kielissä	'language' (32)
vesi	veden	vettä	vesiä	vesissä	'water' (10)
lasi	lasin	lasia	laseja	laseissa	'glass' (4)
nalle	nallen	nallea	nalleja	nalleissa	'teddy' (9)
kirje	kirjeen	kirjettä	kirjeitä	kirjeissä	'letter' (78)

they provide information that the appropriate class is restricted to 4, 9, or 78, but not 8, given that the inessive plural in class 8 is *ovissa*, not *oveissa*. Certain cells among these classes resolve class assignment more reliably than others. For example, *kirjeitä*, the partitive plural of *KIRJE*, appears unique among the forms in the partitive plural column and, therefore, is serviceable as a diagnostic cell for membership in class 78. This becomes particularly clear when we compare this form with the partitive plural forms *laseja* and *nalleja*: even in conjunction with the previously mentioned inessive plurals, these forms do not resolve class assignment between 4 and 9. This is accomplished, however, by comparing the partitive singular forms, *lasia* and *nallea*, or several other contrasts that would serve just as well.

These class-specific sets are reminiscent of the notion of **dynamic** principal parts, which Finkel and Stump (this volume) contrast with what they term “static” and “adaptive” inventories. In fact, there are many equally good alternative sets of principal parts for Finnish, and many more solutions that are almost as good. We speculate that this is a recurrent feature of complex morphological systems (reminiscent of resilience in biological systems). Even though there may be a few very hard cases or true irregulars, in general most cells in the paradigm of most words are of value in predicting the form of most other cells.

As the traditional principal part inventories in Table 3.5 show, the information that facilitates paradigm cell filling in Finnish is not localized in a single form or even in a class-independent set of forms. Instead, forms of an item are partitioned into cohort sets or “subparadigms” that share “recurrent partials.” One pair of subparadigms in Finnish declensions are distinguished by what are conventionally termed the “basic form” and the “inflectional stem” of an item. A typical pattern is illustrated by the paradigm of *ovi* ‘door’, in which the basic form *ovi* realizes the nominative singular and underlies the partitive and inessive plurals, and the inflectional stem *ove* underlies the genitive and partitive singular forms. As in Saami, the organization of cells into subparadigms identifies the form of other declensional cohorts, while variation in the structure of subparadigms across items facilitates the identification of declension classes.

Given this overview of the patterns in Table 3.5, we now outline how to calculate the joint and conditional entropy of the corresponding paradigm cells. Let us first consider how many distinct realizations of the genitive singular are exhibited in Table 3.5. From a traditional perspective, there is exactly one affixal realization, given that “[t]he genitive singular ending is always *-n*, which is added to the inflectional stem” (Karlsson 1999: 91). However, this description already presupposes knowledge of the inflectional stem, which is precisely the type of information that a speaker may need to deduce in order to solve the PCFP. To avoid presupposing information about the organization of Finnish declensions, it is useful to adopt more structurally agnostic descriptions in terms of a “base”, which underlies the basic form (and, usually, the inflectional stem), and an “ending” (which may include the theme vowel of the inflectional stem).¹⁰ On this type of description, the six inflectional classes in Table 3.5 exhibit four distinct realizations. In classes 8, 32, and 9, the genitive singular ends in *-en*. In class 10, it ends in *-en* and the base exhibits a change in the stem consonant. In class 9, it ends in *-in*, and in class 78 it ends in *-een*. If we assume that each of the six declensions has a probability of $\frac{1}{6}$, then the entropy $H(\text{GEN.SG})$ is:

$$\begin{aligned} H(\text{gen.sg}) &= -\left(\frac{3}{6} \log_2 \frac{3}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6} + \frac{1}{6} \log_2 \frac{1}{6}\right) \\ &= 1.792 \end{aligned}$$

Repeating this calculation for each of the cells in the paradigm, we get: The expected entropy $E[H]$ is the average across all cells. Producing a randomly

	Nom Sg	Gen Sg	Part Sg	Part PI	Ines PI	$E[H]$
H	0.918	1.792	2.252	1.459	1.000	1.484

chosen cell of the paradigm of a randomly chosen lexeme (assuming that the declensions are equally likely) requires on average 1.484 bits of information.

Given the paradigms in Table 3.5, we can also calculate the pairwise conditional entropy. Suppose we know that the NOM.SG of a particular lexeme ends in *-i*. What is the genitive singular? Our information about the NOM.SG rules out classes 9 and 78, so we are left choosing among the remaining four classes with three different GEN.SG realizations. Given this information, the uncertainty in the GEN.SG becomes:

¹⁰ This type of pretheoretical description is found particularly in pedagogical grammars and descriptions. For example, noun classes 16–22 in Oinas (2008: 57f.) distinguish *i*- and *e*-stem nouns in terms of the surface variation in their genitive singular forms.

$$\begin{aligned}
 H(\text{GEN.SG}|\text{NOM.SG} = -i) &= -\left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4}\right) \\
 &= 1.5
 \end{aligned}$$

In other words, knowing that the *NOM.SG* ends in *-i* gives us $1.793 - 1.5 = 0.292$ bits of information about the form of the *GEN.SG*. And, if instead we know that the *NOM.SG* of a particular lexeme ends in *-e*, then we must choose between two declensions with two *GEN.SG* realizations, and the entropy is:

$$\begin{aligned}
 H(\text{GEN.SG}|\text{NOM.SG} = -e) &= -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) \\
 &= 1
 \end{aligned}$$

Assuming again that all declensions are equally likely, the probability that the *NOM.SG* of a particular lexeme actually ends in *-i* is $\frac{4}{6}$, and the probability that it ends in *-e* is $\frac{2}{6}$. So, on average, the uncertainty in the *GEN.SG* realization of a lexeme given we know that lexeme's *NOM.SG* realization will be:

$$\begin{aligned}
 H(\text{GEN.SG}|\text{NOM.SG}) &= \frac{4}{6} \times 1.5 + \frac{2}{6} \times 1.0 \\
 &= 1.333
 \end{aligned}$$

In other words, the *NOM.SG* gives us, on average, $1.793 - 1.333 = 0.46$ bits of information about the *GEN.SG*. Table 3.6 gives the pairwise conditional entropy of a column given a row. That is, e.g., $H(\text{NOM.SG}|\text{INES.PL})$ is 0.541 bits.

The row expectation $E[\text{row}]$ is the average conditional entropy of a column given a particular row. This is a measure of the **predictiveness** of a form. By this measure, the partitive singular is the most predictive form: if we know the partitive singular realization for a lexeme and want to produce on other paradigm cells chosen at random, we will require only 0.250 bits of additional information on average. In contrast, given the nominative singular, we would

TABLE 3.6 Conditional entropy $H(\text{col}|\text{row})$ of Finnish *i*-stem and *e*-stem nouns

	Nom Sg	Gen Sg	Part Sg	Part Pl	Ines Pl	$E[\text{row}]$
Nom Sg	—	1.333	1.667	0.874	0.541	1.104
Gen Sg	0.459	—	0.459	0.459	0.459	0.459
Part Sg	0.333	0.000	—	0.333	0.333	0.250
Part Pl	0.333	0.792	1.126	—	0.000	0.563
Ines Pl	0.459	1.252	1.585	0.459	—	0.939
$E[\text{col}]$	0.396	0.844	1.209	0.531	0.333	0.663

need an addition 1.104 bits of information on average. The column expectation $E[\text{col}]$ is the average uncertainty given a row remaining in a particular column.

In contrast to the row expectations, this is a measure of the **predictedness** of a form. By this measure, the inessive plural is the most predicted form: if we want to produce the inessive plural for a lexeme and know some randomly selected other form, we will require on average another 0.333 bits of information.

One cannot of course draw any general conclusions about the implicational structure of Finnish declensions from the calculations in Table 3.6, given that they are based on a small subset of patterns, and that they assume that all classes and variants are equiprobable. Nevertheless, it should be clear that the method applied to this restricted data set scales up, as the description becomes more comprehensive through the addition of further patterns and as it becomes more accurate through the addition of information about type and token frequency.

3.3.3 *Tundra Nenets*

The present section now extends the approach outlined above in order to provide a preliminary case study of nominal inflection in Tundra Nenets (Samoyed branch of Uralic). The basic question is this: Given any Tundra Nenets inflected nominal word form, what are the remaining 209 forms of this lexeme for the allowable morphosyntactic feature property combinations CASE: {nom, acc, gen, dat, loc, abl, pro}, NUMBER: {singular, dual, plural}, POSSESSOR: {3 persons \times 3 numbers}? The problem can be schematized as in (5a) and (5b). Specifically, given exposure to a stimulus such as that in (5a), the nominal *nganu' mana* 'boat (plural prosecutive)', what leads to the inference that its nominative singular form is the target *ngano*? In contrast, if confronted with the plural prosecutive of the nominal *wíngo' mana* 'tundra (plural prosecutive)', what leads to the inference that its nominative singular is the target *wí*?

- | | | | | | | | |
|-----|----|--------------------|--------------|----|----|--------------------|---------------|
| (5) | a. | Stimulus: | Target | vs | b. | Stimulus | Target |
| | | <i>nganu' mana</i> | <i>ngano</i> | | | <i>wíngo' mana</i> | <i>wí</i> |
| | | boat.PL.PROS | boat.SG.NOM | | | tundra.PL.PROS | tundra.SG.NOM |

In line with the hypotheses set out in the previous section, we must identify the patterns of interpredictability for a subset of Tundra Nenets nominal declensions within and across subparadigms. This entails stating the principles of arrangement within and across stem types. For the absolute declension (i.e., nonpossessive, nonpredicative nominals), lexical categories are divisible into the gross stem-type classification in Table 3.7 (again ignoring the role of syllabicity;

TABLE 3.7 Tundra Nenets nominal types (Salminen 1997, 1998)

Type 1 (T1):	stem ends in C (other than a glottal stop) or V;
Type 2 (T2):	subtype 1: stem ends in nasalizing/voicing glottal (')
	subtype 2: stem ends in non-nasalizing/devoicing glottal ('')

see Salminen (1997, 1998) for a detailed exposition of types).¹¹ For simplicity, we demonstrate the basic pattern with an example of Type I in Table 3.8.

Examination of Table 3.8 yields a basic observation: the nominal paradigms for all stem classes are partitioned into subparadigms, each of which is defined by the presence of a characteristic and recurring stem (*ngano*, *nganu*, or *nganoxo*). In what follows we will refer to these forms as recurrent partials and the sets in which they recur as coalitions or alliances (or cohorts) of forms. This brings out the following generalization about Tundra Nenets absolute nominal paradigms:

Subparadigms are domains of interpredictability among alliances of word forms, rather than sets of forms derived from a single base.¹²

An approach based on recurrent partials, and patterns of relatedness among forms, develops the approach in Bochner (1993), in which no form need serve as a privileged base form among different surface expression of a lexeme.

TABLE 3.8 Type I: Polysyllabic vowel stem: *ngano* 'boat'

	Singular	Plural	Dual
Nominative	<i>ngano</i>	<i>ngano''</i>	<i>nganoxo'</i>
Accusative	<i>nganom'</i>	<i>nganu</i>	<i>nganoxo'</i>
Genitive	<i>ngano'</i>	<i>nganu''</i>	<i>nganoxo'</i>
Dative-Directional	<i>nganon'</i>	<i>nganoxo''</i>	<i>nganoxo' nya'</i>
Locative-Instrumental	<i>nganoxona</i>	<i>nganoxo'' na</i>	<i>nganoxo' nyana</i>
Ablative	<i>nganoxod</i>	<i>nganoxot</i>	<i>nganoxo' nyad</i>
Prolative	<i>nganowna</i>	<i>nganu'' mana</i>	<i>nganoxo' nyamna</i>

¹¹ There are phonological properties associated with particular glottal-final stems (as in Saami and Finnish) that decrease the uncertainty of predicting class assignment and related forms of words within the class. For example, the occurrence of a specific allomorph, e.g., *wingana* (where *-gana* is part of a family allomorphs such as *-xana* and *-kana*) leads to the inference that this word belongs to the class of stem-final nasalizing glottals. In this way, surface allomorphy can be used as a diagnostic clue for guiding paradigm-based inferences.

¹² As expected from positing that patterns of inflected forms exist, there is a need to access certain of them for purposes of derivational relatedness in Tundra Nenets. In particular, there are at least two verbal derivation operations built upon the form used to express genitive plural nominals. See Kupryanova (1985: 139).

Regardless of whether a stem exists as an independent word, all these systems share the property that they have clusters of related forms where it is at least somewhat arbitrary to take any one form as basic. This is what I take to be defining characteristic of a paradigm. Thus, we need a way to relate to the various members of paradigm directly to each other without singling out any one of them as a base for the others. (Bochner 1993: 122)

On this type of analysis, alliances of word forms share recurrent partials, but the elements in such alliances need not be thought of as bearing derivational or “constructive” relations (in the sense of J. P. Blevins 2006*b*) to one another, let alone to a single isolable base form. The relations among members of subparadigms are symmetrical, since there is no one form that serves as the base from which the others are derived.¹³ This organization is depicted in Figure 3.1, which partitions the Tundra Nenets nominal declension into three alliances of forms. Each form in a subparadigm provides information about other forms in the same subparadigm. The members of a subparadigm share partials, thereby making an alliance a system of interpredictability among related word forms.¹⁴

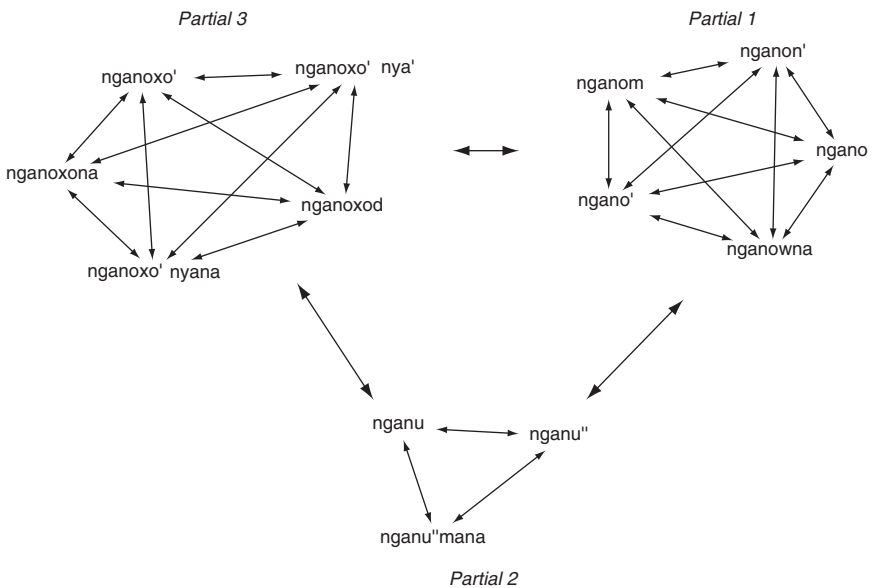


FIGURE 3.1 Symmetric paradigm organization

¹³ However, the lack of a single privileged base does not entail that there cannot be multiple subparadigms in which a particular recurring form (a partial) serves a pivotal role.

¹⁴ This is compatible with Albright’s observation that “when we look at larger paradigms . . . it often appears that we need local bases for each sub-paradigm (something like the traditional idea of principal parts, or multiple stems)” (Albright 2002*a*: 118).

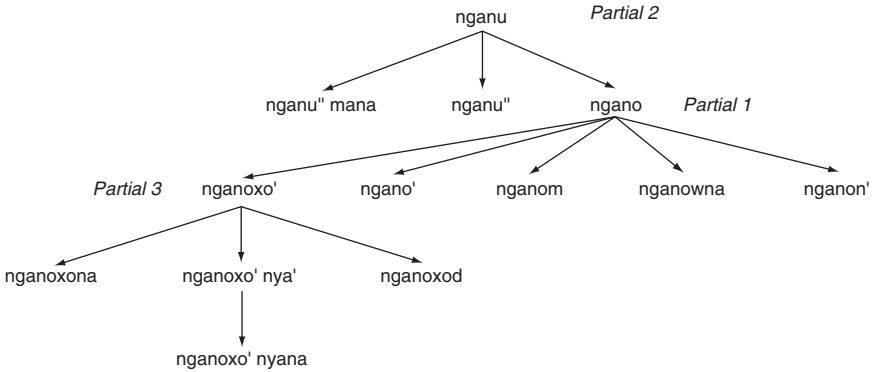


FIGURE 3.2 Asymmetric paradigm organization

In contrast, derivational or constructive relations based on a single form are asymmetric in assuming that some specific form is predictive of the other forms. An asymmetric structure, organized in terms of local bases, is depicted in Figure 3.2. In contrast to Figure 3.1, each subparadigm contains a base from which the rest of the forms in it are derived. There is no notion of interpredictability of the sort manifest in Figure 3.1: the base gives information about derived forms, but the derived forms need not give information about a base.

3.3.3.1 Implications across subparadigms The strategy we have chosen to demonstrate the utility of symmetric organization is to focus on the most challenging and problematic instance of relatedness between two word forms within Tundra Nenets nominal paradigms, specifically the *NOM.SG* and *ACC.PL*. The logic of this task is straightforward: if we can identify a direction with reliably low conditional entropy, i.e., most predictive, between the two least transparently related word forms, then there is reason to believe that asymmetric derivation may be viable. In other words, one could hypothesize that knowing e.g., *NOM.SG*, would suffice to predict the *ACC.PL* across all classes, either directly, or by identifying a common base that underlies both forms. In contrast, the symmetric proposal is compatible with a situation in which there is no single reliably predictive form, but that classes are organized into patterns of interpredictability within alliances of forms.

Consider the pairs of *NOM.SG* and *ACC.PL* forms in Table 3.9. A comparison of the forms in the columns reveals that there is indeterminacy or uncertainty with respect to predictability in both directions. For example, while the *ACC.PL* of ‘boat’ and ‘harnessed deer’ both end in the vowel *-u*, their *NOM.SG* forms

TABLE 3.9 Tundra Nenets inflected nominals

Nom Sg	Acc Pl	
ngano	nganu	'boat'
lyabtu	lyabtu	'harnessed deer'
ngum	nguwu	'grass'
xa	xawo	'ear'
nyum	nyubye	'name'
yí	yíbye	'wit'
myir	myirye	'ware'
wí'	wíngo	'tundra'
we'	weno	'dog'
nguda	ngudyi	'hand'
xoba	xob	'fur'
sawənye	sawənyi	'magpie'
tyírtya	tyírtya	'bird'

end in *-o* and *-u* respectively. Likewise, while the NOM.SG of 'boat' ends in *-o*, the ACC.PL of 'grass' ends in *-o* and its NOM.SG ends in the consonant *-m*.

The basic question is, given exposure to one form, how well can one predict the other? This is just the PCEP relativized to Tundra Nenets. In the following preliminary study, we use data from a corpus of 4,334 nominals. These are extracted from Salminen's compilation of 16,403 entries, which is based on Tereshchenko's Nenets-Russian dictionary (1965/2003). The compilation specifies meaning, frequency, as well as the stem-class assignment. We explore the relative predictiveness of NOM.SG and ACC.PL, with the following query in mind: which of these forms, if either, is more useful for predicting the other? The first calculation maintains the idealization adopted in the analyses of Saami and Finnish and assumes that all declension classes are equally likely. We start by identifying 24 different types of nominative singulars. The entropy of this distribution is $H(\text{NOM.SG}) = 4.173$ bits. There are likewise 29 different types of accusative plurals, and their entropy is $H(\text{ACC.PL}) = 4.693$ bits. Taken together, there are 43 nominal 'declensions' represented in the compilation (each declension being a combination of a NOM.SG realization and an ACC.PL realization), and the joint entropy of the two forms is $\log_2 43 = 5.426$ bits.

These calculations assume (as in the case of Saami and Finnish) that all declensions are equally likely. However, it is clear from the compilation that all declensions are **not** equally likely. In fact, the distribution of type frequencies across declensions is highly skewed: the five most frequent declensions account for more than half of the noun lexemes (see Figure 3.3 for the complete distribution). Taking the type frequencies of declensions into

account, we now find that the entropy associated with each individual form is $H(\text{NOM.SG}) = 3.224$ bits and $H(\text{acc.pl}) = 3.375$ bits. The true joint entropy $H(\text{NOM.SG}, \text{ACC.PL})$ is 3.905 bits, a level of uncertainty equivalent to 15 equiprobable declensions.

Having quantified the degree of uncertainty in the choice of *NOM.SG* and *ACC.PL* types individually, we can now calculate predictability of one realization given the other, using conditional entropy $H(Y|X)$. Consider first the task of predicting the *ACC.PL* form from the *NOM.SG*. We can evaluate the difficulty of this prediction using the conditional entropy $H(\text{ACC.PL}|\text{NOM.SG})$, the uncertainty in the *ACC.PL* given the *NOM.SG*. Out of the $24 \times 29 = 696$ possible pairings of *NOM.SG* and *ACC.PL* types, 43 are actually attested in the lexicon. In some cases, knowing the *NOM.SG* of a word uniquely identifies its *ACC.PL*, e.g. a word ending in *-ye* in the *NOM.SG* always has an *ACC.PL* in *-yi*. For such words, once we know the *NOM.SG* there is no uncertainty in the *ACC.PL* and the conditional entropy $H(\text{ACC.PL}|-ye) = 0$ bits. In other cases, however, knowing the *NOM.SG* narrows down the choices for the *ACC.PL* but does not uniquely identify it. For example, polysyllabic words whose *NOM.SG* ends in *-ya* might have an accusative plural in *-∅*, *-yi*, or *-e*. Furthermore, of the 289 polysyllabic lexemes with a *NOM.SG* in *-ya*, 268 have an *ACC.PL* in *-yi*, 19 in *-∅*, and only 2 in *-e*. So, the entropy is:

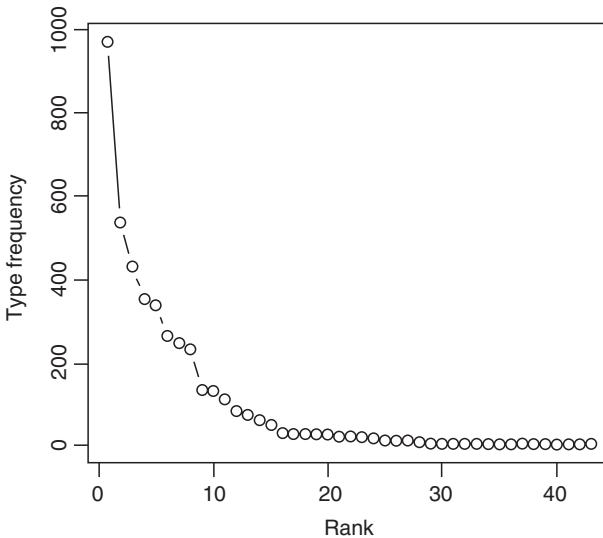


FIGURE 3.3 Type frequencies of Tundra Nenets nominal declensions, by rank

$$\begin{aligned}
 H(\text{ACC.PL}|\text{NOM.SG} = -ya) &= -\left(\frac{268}{289} \log_2 \frac{268}{289} + \frac{19}{289} \log_2 \frac{19}{289} + \frac{2}{289} \log_2 \frac{2}{289}\right) \\
 &= 0.410 \text{ bits}
 \end{aligned}$$

Averaging across the whole (sample) lexicon, the uncertainty in the ACC.PL given the NOM.SG is $H(\text{ACC.PL}|\text{NOM.SG}) = 0.681$ bits. In other words, the NOM.SG “predicts” all but 0.681 of the 3.375 bits of uncertainty previously calculated for the ACC.PL. Now, if we switch directions, going from ACC.PL to NOM.SG, it turns out that the conditional entropy $H(\text{NOM.SG}|\text{ACC.PL}) = 0.530$. In other words, the ACC.PL “predicts” all but 0.530 of the 3.224 bits in the NOM.SG. Since the conditional entropy is closer to 0 in the latter than in the former, the ACC.PL appears to be more helpful for predicting the NOM.SG than vice versa, but only by a slim margin. More importantly, neither conditional entropy is 0 bits or close to it, meaning neither form is especially useful for predicting the other.

Hence, there is no principled grounds for hypothesizing that one form or the other serves as (or even identifies) a single privileged base. Either choice would still leave a large inventory of irregular pairings to be memorized by the language learner. This arbitrary choice is avoided on a symmetric account, where there is no need to suppose that some forms are reliably predictable from others. Instead, a symmetrical proposal posits alliances which cohere into coalitions of interpredictable forms and which together partition the entire paradigm. We do not expect forms that take part in different alliances to be mutually predictive, so the fact that knowledge of a member of one alliance does not reliably reduce uncertainty about a member of another is not surprising.

More positively, the utility of alliances becomes clearer if one considers the distribution of Tundra Nenets forms. Although the NOM.SG and ACC.PL are equally unsuitable as single bases, the NOM.SG will still make a more prominent contribution to defining the implicational structure of a paradigm, given that speakers are far more likely to encounter the NOM.SG form of a noun than the ACC.PL form. The distributional difference between these forms is reflected in the frequency counts in Table 3.10, representing the 12,152 noun tokens in Salminen’s sample sentence corpus. The NOM.SG represents 33.8 percent of the tokens, while the ACC.PL represents only 2.7 percent. Speakers cannot just assume that the most frequent form is the most useful for solving the PCFP, given that the NOM.SG is not even a reliable predictor of the ACC.PL. The ACC.PL itself is an even less suitable candidate. Even if the predictive value of the ACC.PL made it potentially useful as a base, the attested frequencies suggest that speakers would have a low likelihood of encountering this form for any

TABLE 3.10 Word-form frequencies in Tundra Nenets

	Singular	Plural	Dual
Nominative	4,117	770	7
Accusative	1,077	355	6
Genitive	3,002	376	5
Dative-Directional	762	89	0
Locative-Instrumental	724	108	0
Ablative	291	50	0
Prolative	372	41	0

given item. The situation is worse yet for forms such as the direct case dual forms, which account for 0.1 percent of the tokens. In fact, no individual word form (other than the *NOM.SG* and the *GEN.SG*) occurs with high enough frequency to be a reliable source of information about a word's inflectional class. This makes Tundra Nenets a challenging language from a "single base" point of view, as speakers cannot be sure of encountering the diagnostic forms necessary to identify a word's inflection class.

However, the issue takes on a different complexion when we look at forms in terms of alliances, organized around the Partial 1, 2 and 3 in Figure 3.1. Although the *ACC.PL* is a relatively low-frequency form, it is predictable from other forms that it is transparently related to. For example, the *GEN.PL* adds a final glottal stop to the *ACC.PL*, as illustrated by the relation between *nganu*, the *ACC.PL* form of 'boat', and the corresponding *GEN.PL nganu''*. Hence, while there is a low likelihood of encountering the *ACC.PL*, there is a much higher likelihood of encountering the **partial** associated with *ACC.PL* (from which the *ACC.PL* can be defined), if paradigms are organized into alliances of interpredictable forms that "pool" the frequency of individual forms. The effect of this structure is shown by the contrast between the form frequencies in Figure 3.1 and the totals in Table 3.11, which sum the token frequencies of all absolute and possessive forms.

The organization of forms into subparadigms thus serves two related functions. On the one hand, high-frequency forms such as the *NOM.SG* or *GEN.SG* identify the shape of lower-frequency members of the same alliance, such as the prolative singular. On the other hand, "pooling" the frequencies of the members of each alliance allows Partial 2, and the forms based on this partial, to be identified either by the *ACC.PL* and the *GEN.PL*, while Partial 3, and forms based on it, can be identified by the locative-instrumental forms or by the ablative singular. By relying on alliances of related forms within subparadigms, speakers may gain reliable cues about the shape of even very low-frequency word forms.

Significantly, accounts that assume an asymmetrical relation between a privileged base and derived forms have no obvious analogue to alliances of mutually reinforcing forms. On such asymmetrical approaches, the patterns within subparadigms appear epiphenomenal, not, as suggested here, as central to the organization of the declensional system and critical to the solution of the PCFP.

3.3.3.2 *Summary* The preceding sections suggest that neither the *NOM.SG* nor *ACC.PL* form can serve reliably as the single base from which the other is predicted. Yet the fact that neither form is fully predictive does not mean that they are uninformative. Instead, the association of forms with subparadigms allows speakers to exploit the fact that partials appear with much higher frequency than any given word-form. Hence, there is no need to encounter a privileged member of an alliance in order to predict allied forms. What is important is just that each alliance contain at least some high-frequency forms and that the aggregate frequency of partials within the alliance is high enough to be useful. In this way, the organization of the Nenets declensional system makes available many of the basic ingredients for a solution to the paradigm cell filling problem.

3.4 Conclusions

We conclude by returning to the questions in (6), repeated from (1), which concern issues raised by traditional solutions to the Paradigm Cell Filling Problem.

- (6) a. What is the structure of units that license implicative relations?
- b. How are units organized into larger structures within a system?
- c. How can one measure implicative relations between these units?
- d. How might the implicative organization of a system contribute to licensing inferences that solve the paradigm cell filling problem?
- e. How does this organization, and the surface inferences it licenses, contribute to the robustness and learnability of complex systems?

This chapter has focused primarily on questions (6a), (6b), and (6c). The central hypothesis has been that words are organized into paradigms and that information-theoretic measures provide an insightful measure of relatedness among members of declension classes. In fact, distinctive patterns of relatedness clearly enter into what it means to be a declension class, with some forms or combinations of forms being more diagnostic of class membership than others. Once conditional entropies for families of forms are identified, they can be used, along the lines we have suggested, to solve the Paradigm Cell

Filling Problem. Individual forms or alliances of forms serve as cues for simplifying the assignment of class membership for novel words on the basis of the analogies provided by the patterns of known words. It is worth emphasizing that the answers to questions (6a), (6b), and (6c) presuppose access to (patterns of) surface word forms. There are, accordingly, several theoretical consequences associated with our results.

First, there must be more to morphological analysis and morphological theory than the distillation of rules or patterns for the composition of individual word forms. In focusing exclusively interest on the syntagmatic dimension of morphological analysis, the post-Bloomfieldian tradition has been led to adopt questionable claims about the nature of the grammatical system and the mental lexicon. Work within this tradition has assumed that morphological analysis consists of identifying morphemes and stating rules for describe morpheme combinations. Larger structures such as words and paradigms tend to be treated as derivative or even as epiphenomenal. The emphasis on identifying minimal units has also fostered the *a priori* belief that the lexicon consists entirely of minimal elements, and, in particular, that productive and regular word forms are not part of the mental lexicon of a speaker, on the grounds that such forms would be “redundant” if they could be constructed from available morphemes and combinatoric rules. Yet a range of psycholinguistic studies has shown that the processing of a given word may be influenced (whether facilitated or inhibited) by other related forms in a way that suggests that the related words are available as elements of a speaker’s mental lexicon (Baayen *et al.* 1997; Schreuder & Baayen 1997; Hay 2001; de Jong 2002; Moscoso del Prado Martín 2003). Another group of studies provide evidence for various types of paradigm-based organization (Baayen & Moscoso del Prado Martín 2005; Milin *et al.* 2009).

The traditional word and paradigm assumptions adopted here appear to be more compatible with these results than the post-Bloomfieldian assumptions that still guide modern generative accounts. In order to unify these perspectives, one might take them to adopt have complementary foci, with WP approaches focusing on whole words and their organization into paradigms, and morphemic accounts focusing on the internal structure and construction of word forms. We suggest that this is misleading. For the languages discussed above and others of comparable complexity, the answer to question (6a) must appeal to the whole words and larger paradigmatic structures recognized in WP approaches. There is little evidence that syntagmatic approaches have any means of characterizing the role that whole words play in morphology, let alone the place of larger paradigmatic structures. In contrast, a WP approach is largely agnostic about the internal structure of complex words. A WP

approach is compatible with an agglutinative **morphotactic** analysis, in cases where such an analysis is motivated. But a WP account is also able to characterize the extraordinary variety of strategies for the creation of complex word forms attested cross-linguistically, without reducing them to an underlying basic structure. In order to arrive at a general answer to question (6a), we suggested that complex words are recombinant gestalts. On this pattern-based view, agglutination is just a particularly simple pattern. Finally, with respect to question (6e), we suggest that it is the very pattern-based nature of morphology – both at the level of individual (types of words) and in their organization into paradigms – that makes even highly complex morphological systems learnable and, by hypothesis, guides the development, maintenance, and change of these systems.

Resolving pattern conflict: Variation and selection in phonology and morphology

Andrew Wedel

4.1 Introduction

Every language system comprises many intersecting levels of organization, each with its own structures and patterns. When these levels overlap, patterns at different levels can come into conflict. For example, phonological regularity may entail morphological irregularity, as when addition of an affix requires a change in a stem. In Catalan, for example, some verbal suffixes are underlyingly stressed such that they may induce a shift in stress in the stem. This interacts with phonological vowel reduction processes in Catalan to result in differences in stem vowel realizations between members of a verbal paradigm, as exemplified below (Wheeler 2005).

- (1) a. /don/ [dónu] [duném] ‘give’
b. /pas/ [pásu] [pəsém] ‘pass’

Morphological regularity in turn can entail phonological irregularity, as when a stem fails to undergo an otherwise regular phonological change upon affixation resulting in the maintenance of consistency across the paradigm. A classic example of such a “paradigm uniformity effect” was noted by Chomsky and Halle (1968) in the occasional absence of an otherwise expected vowel reduction to schwa in English. For example, the words ‘comp[ə]nsation’ and ‘cond[ɛ]nsation’ have very similar prosody, but the latter maintains a full vowel pronunciation of [ɛ] rather than the expected reduction to schwa that we see in ‘comp[ə]nsation’. On the basis of this and a number of other similar cases, Chomsky and Halle argue that the phonological irregularity of

cond[ɛ]nsation arises because it is constrained to remain similar to its base of affixation: compare the associated bases ‘comp[ə]nsate’ and ‘cond[ɛ]nse’.

Another example involving stress can be found in Polish (Rubach and Booij 1985). Polish has primary stress on the penultimate syllable, while preceding syllables are organized into left-headed feet aligned to the beginning of the prosodic word (compare example 2a and b). When prefixed with an enclitic, initial stress shifts such that the prosodic word begins with a foot. However the remainder of the foot structure of the stem remains parallel to the form without the enclitic in violation of the default stress pattern (compare Figures 2c and d).

- | | | | |
|--------|-------------------------|--|---------------------------------------|
| (2) a. | kònstàntỳnopòlitànczyk | ($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$) | ‘Inhabitant of
Constantinople-NOM’ |
| b. | kònstàntỳnopòlitanczýka | ($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$) σ ($\acute{\sigma}\sigma$) | ‘Inhabitant of
Constantinople-GEN’ |
| c. | àmerykànía | ($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$) | ‘American-GEN’ |
| d. | dò amerykànía | ($\acute{\sigma}\sigma$) σ ($\acute{\sigma}\sigma$)($\acute{\sigma}\sigma$) | ‘to an American-GEN’ |

Analogy, in the sense of pattern extension, is a significant route for change in systems of categories (Itkonen 2005). This chapter is an exploration of the ways that conflicting patterns at different levels of organization may mutually influence one another to produce change. Working within an evolutionary framework (see, e.g., Blevins 2004; Pierrehumbert 2006; Croft 2008), I have argued that similarity-biased variation can contribute to the entrenchment and extension of regular patterns over many cycles of language use and transmission (Wedel 2007). An evolutionary approach to pattern development and change is supported by the great deal of evidence that lexical memory is richly detailed at a number of levels, rather than limited to storage of symbolic, contrastive features as proposed in many classical models (reviewed in Pierrehumbert 2003). Within a model incorporating this evidence for rich memory, biases in production and perception toward previously experienced forms create a positive feedback loop promoting pattern entrenchment (Wedel 2006, 2007, reviewed in Pierrehumbert 2006). Given that a given system can potentially evolve toward many different meta-stable states, a task for anyone working within this evolutionary model of language pattern development is to understand what factors encourage or inhibit the transition from a given pattern into another. Recent examples of work in this area can be found in Blevins (2004), Mielke (2004), Chitoran and Hualde (2007), J. Blevins (2008), and many others. In this chapter I argue that pattern conflict across distinct levels of organization can be understood in a feedback-driven model of change as an instance of multilevel selection, and that this can

help us think productively about the role of within-category variance in promoting or inhibiting change throughout the language system.

In the following section I review the role of noise in creating similarity biases in category processing. In Section 4.3 I go over some of the kinds of language change in which similarity-biased error may plausibly play a role. In Section 4.4 I review how variation introduced by error influences the development of patterns within a rich memory model, as well as the use of evolutionary theory to model this process. Section 4.5 discusses possible mechanisms for similarity biases in production and perception that can feed language change. Section 4.6 introduces multilevel selection as a potentially useful way to think about conflicts between different levels of generalization. Finally, Section 4.7 presents an illustrative simulation of a multilevel selection at work in a model lexical system evolving under competing attractors formed by distinct phonological and morphological regularities.

4.2 Error and similarity bias in categorization

Information processing is always errorful to some degree due to noise. The simplest error pattern arises in processing of individual bits of information in which there are just two possible states, e.g., 0 and 1. In this case, noise can only result in the transformation of one bit value into the other.

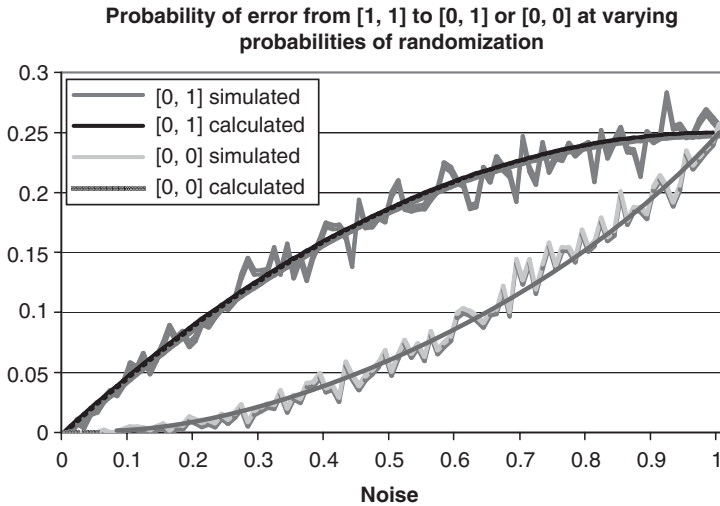
$$(3) \{ \dots, 1, 1, 1, \dots \} \rightarrow \text{noise} \rightarrow \{ \dots, 1, 0, 1, \dots \}$$

Much of the processing in language, however, involves processing compositional signals in which the unit of interest is above the level of an indivisible bit. For example, a word is composed of subsidiary units of information, such as segments. Successful transmission of a higher-order category such as this requires that both the information source and target have access to a common lower-level information pattern identifying the category, e.g., a segment sequence. In this case, there are two possible outcomes of noise in processing. If noise results in a pattern that is not successfully matched to any existing category, processing fails altogether at that level. However, noise can also result in a match to a different category, as when an American English speaker utters *can't* but I understand *can*.

In a system in which categories can overlap to varying degrees, that is, can share variable amounts of lower-level information, noise will always favor mismatches between more similar over less similar categories. As an example, consider four categories each comprising two bits of information: $\{[1, 1], [1, 0], [0, 1], [0, 0]\}$. At any level of noise below that producing complete

randomization, the odds that $[1, 1]$ will be mismatched to $[0, 1]$ or $[1, 0]$ is always greater than the probability of matching to $[0, 0]$. As an illustration, (4) shows the rate of matching $[1, 1]$ to $[0, 1]$, and to $[0, 0]$ respectively, at varying noise levels, where a noise level of 1 represents complete randomization of original information. Numerically predicted rates are shown as well as simulated rates averaged over 1000 trials at each noise level.

- (4) Pattern-matching error to similar versus less similar categories under noise.



Given that language involves the processing of compositional categories that vary in their similarity along various dimensions, noise-driven mismatch errors will always be biased toward similar categories. In previous work I have argued that similarity-biased, “analogical” error can serve as a seed for phonological change and entrenchment of patterns within a rich-memory model of language production and perception (Wedel 2004, 2006, 2007). Here, I will explore some consequences of the hypothesis that a general similarity-biased error also contributes to analogical change at the morphological level. Because dimensions of phonological and morphological similarity can cut across one another, similarity-biased error and variation should set up conflicts between these distinct kinds of regularity. My goal in the next sections is to show that considering analogical change of all kinds to be initiated by similarity-biased error has the potential to shed light on the outcomes of conflict between and among phonological and morphological regularities (Sturtevant 1947).

4.3 Pattern extension in phonology and morphology

Many sound changes are “unnatural” in the sense that they do not appear to originate in common articulatory or perceptual tendencies. Some of these appear instead to originate in pattern extension. For example, in phonology, sound patterns can be extended from an original, “natural” context into contexts in which the change is not clearly phonetically motivated (for examples, see Mielke 2004: 102–14; J. Blevins 2006a).

In morphology, both *leveling* and *extension* changes can be considered instances of pattern extension (discussed in Deutscher 2001; Hock 2003). In leveling, members *within* a paradigm become more alike in some way. For example, the historical stem-final [f ~ v] alternation in the singular-plural pair *dwarf* ~ *dwarves* has leveled for many speakers of American English to *dwarf* ~ *dwarfs*. Paradigmatic extension occurs when a change creates a relationship within one paradigm that is parallel in some way to a relationship holding in another. For example, the originally regular present-past paradigm of *dive* ~ *dived* has shifted for many speakers to *dive* ~ *dove*, presumably by extension on the model of the group containing *drive* ~ *drove*, *ride* ~ *rode*, etc.

None of these phonological or morphological patterns can be fully understood without making reference to the existing language system. Given that learners and adult users alike have some knowledge of the ambient linguistic system, there are two conceptually distinct pathways by which patterns in the existing system can influence change: (i) by influencing the range of variants presented by adults to learners as input, and (ii) by influencing the ways that learners organize this input as they bootstrap between input and their current system toward the adult system (Pierrehumbert 2003; cf. CHANCE and CHOICE in the framework of Evolutionary Phonology, Blevins 2004¹). In both cases, similarity biases can accentuate asymmetries within the experience of an individual. Within a rich-memory model of language production and processing (e.g., Pierrehumbert 2001; Bybee 2002; Wedel 2004, 2007), this asymmetry in experience is recorded in a corresponding asymmetry in the language system at some level. What dimensions of similarity are most salient in a particular system is an empirical question, dependent on both relatively universal as well as system-specific details (see e.g., Albright (this volume) and Pierrehumbert (2006) for discussion of these issues).

¹ In Evolutionary Phonology (Blevins 2004, 2006b) CHANCE is a pathway of change by which features of a percept are intrinsically ambiguous, allowing different learners to impose different underlying structure on a common surface form. CHOICE is an abstractly similar pathway for differential development of underlying structure, where given a range of variant productions a single category one learner chooses a different prototypical form than another.

4.4 Rich memory, feedback, and evolution

There is abundant evidence that the mental lexicon stores a great deal of information about perceived variants of lexical forms. In turn, there is evidence that new experiences continually contribute to this store of information, and that this information biases both subsequent perception (e.g., Johnson 1997; Guenther *et al.* 2004; Eisner and McQueen 2005) and production (Goldinger 2000; Harrington *et al.* 2000). As a result, processing a particular instance of a form increases the probability that a similar form will be processed in the same way in the future, and that corresponding forms will be produced in a similar way in the future. This creates positive feedback that promotes the entrenchment of patterns over many cycles of production and perception in acquisition, and to some degree in adult usage as well (Wedel 2006, 2007, reviewed in Pierrehumbert 2006).

In a clever demonstration of feedback between perception and production Goldinger (2000) had a group of subjects produce a baseline recording of a set of words. The next day they *heard* the same words spoken some number of times in a particular voice. They returned five days later and were recorded again reading the same list of words. For each word recorded by each subject, an AXB test stimulus was made from (i) the subject baseline recording of the word, (ii) the word as heard by the subject the second day, and (iii) the final subject test recording of that word (where the order of the baseline and test recording was random). These recordings were played for a separate group of listeners who were given the task of rating which of the two subject recordings of the word was more like the middle reference recording in the other voice. The listeners identified the second test recording as more similar to the reference recording at significantly above chance, indicating that for the subjects, phonetic details of a pronunciation heard five days earlier had a significant influence on their own current pronunciation of that word.

Within a model of language in which variation within and across categories can be stored in some form, reproduced, and transmitted, the system as a whole can change through evolutionary processes (e.g., Wedel 2006; Kirby 2007; Croft 2008). The most well-known mechanism for a reproducing population to evolve over time is through selection, in which some variant elements in the population reproduce more than others via some interaction within the system. As long as there is some mechanism for variation to arise and persist, selection can favor some variants over others in some way, thereby altering the distribution of characteristics within the population over time. Although some rich-memory models assume that the only content of

categories is in the form of fully detailed exemplars (reviewed in Tenpenny 1995), there is evidence that behavior also proceeds through use of independent, more abstract generalizations about input data (e.g., Kuehne *et al.* 2000; Albright this volume). For the purposes of the argument here, provided that within-category variation can persist in the system at some level – whether at the level of exemplars or of generalizations about some form – selection among these variants can result in change in the system.

When patterns conflict within systems including positive feedback, the most stable outcome is dominance of one pattern over the other. Examples from familiar life include the direction that a ball rolls down a symmetrical hill starting from the top. At the top all directions may be equally likely, but once the ball begins moving in a particular direction, other directions become increasingly unlikely. A more complex example comes from economics, where in many cases the larger a company is, the better it can compete. All else being equal, in this situation the most stable state may be a monopoly (Sharkey 1982).

In previous work, I have argued that similarity-biased errors in production and perception may serve as an underlying cause of the development of regular phonological patterns in language through positive feedback, despite the ability of the language system to store and use otherwise predictable information (Wedel 2007). The development of consistent patterns in morphology has also been argued to arise through positive feedback over many cycles of errorful learning (e.g., Hare and Elman 1995; Kirby 2001). Hare and Elman, for example, showed that sequential errorful learning and production by connectionist networks could reproduce the general trajectory of pattern changes that occurred in present-past verb paradigms between two stages of Old English. They first trained a connectionist network to reproduce a large set of Old English present-past verb-form pairs, and then used the output of this network as the learning input to a subsequent naïve network, the output of which served as input to the next, and so on. Because errors in network outputs tend to favor robust generalizations at the expense of less well-attested patterns, the result over many transfers was a gently accelerating consolidation as, for example, the incipient “regular” past-tense pattern became increasingly robust within the data.

4.5 Similarity biases in production and perception

In any process that distinguishes between categories, the rate of error in element identification or manipulation due to noise will be greater between more similar categories relative to less similar categories. Processes in language use that provide opportunities for these kinds of similarity-biased errors include (i) motor entrenchment in production, (ii) the magnet effect in perception, and

(iii) the application of relational categories to compose related forms. Motor entrenchment is a general property of motor systems in which practiced motor routines bias future motor execution in some relation to similarity (Zanone and Kelso 1997). This sets up a positive feedback loop in which, *ceteris paribus*, less frequent production variants should be steadily deformed toward more frequent production variants over time (Bybee 2002, discussed in Wedel 2006, 2007). On the perceptual side, the perceptual magnet effect (Kuhl 1991, 1995) provides another potential source of positive feedback which can act to enhance the similarity of forms over time. The perceptual magnet effect refers to the finding that percepts tend to be biased systematically toward the centers of categories relative to the stimuli that gave rise to them. This systematic warping should pull similar pronunciations closer together over time through feedback between perception and production (Wedel 2007).

Both motor gestures and linguistically relevant sound categories often have a relational internal structure, meaning that they cannot be fully characterized by a simple list of properties. Instead, these categories must include some higher-order relational information. Phonological examples with concrete internal relational structure include sound-categories with temporally ordered gestures such as diphthongs, affricates, and contour tones. The central importance for language of such “relational categories” has been discussed at length by Dedre Gentner and colleagues in the context of semantics (Gentner and Kurtz 2005). Morphophonological patterns are also relational, in that they describe some mapping between forms (Bybee 1985). These patterns are often described in terms of rules, but they may be described as well in terms of relational categories, identified with, for example, the large number of possible patterns in the relationship between present and past forms of English verbs (Albright and Hayes 2002). Generalizations (whether expressed as rules or relational categories) play a role in production or identification of linguistic forms whenever some form is reconstructed through reference to some other form or pattern. This is analogy. The parade example of this use is in the production of a novel form fitting a pattern. In this case, a large body of research indicates that the applicability of a generalization to a novel form is gradient and dependent on similarity to other forms that are covered under that generalization (e.g., Long and Almor 2000; Albright 2002a; Krott *et al.* 2002; Ernestus and Baayen 2003). For example, the novel present-past verb form pair ‘spling ~ splung’ is highly similar to members of a significant subpattern in English verb forms including ‘sing ~ sung’, ‘spring ~ sprung’, etc. Despite the fact that ‘spling’ is a novel form, ‘splung’ is rated as a very good possible past-tense form for this verb (Albright 2002a), contrary to models that assume that all novel forms will be produced via a default pattern (e.g., Pinker 1991).

It has been noted that production of previously learned, morphologically complex forms within a paradigm might proceed by direct retrieval from memory, or through reconstruction from a base or related word-form using an associated generalization (e.g., Baayen 1992; Schreuder and Baayen 1995; Alegre and Gordon 1999). Error in application of a generalization in this process can result in an extended or leveled output pattern depending on the source of the generalization (Hock 2003). Extension of a compositional pattern results in a leveled output, as when speakers of English occasionally produce the past tense of an irregular verb regularly, e.g., ‘teached’ rather than ‘taught’. Conversely, extension of irregular patterns also occur, and have been shown to be more likely in bases that share phonological features with the set of forms exhibiting that irregular pattern (Bybee and Modor 1983; Long and Almor 2000; Albright and Hayes 2002), as for example when the past tense of ‘bring’ is produced as ‘brang’ by analogy to the ‘sing ~ sang’ group of verbs.

There are a wide variety of generalizations that are potentially involved in production and perception of any linguistic form, from lower-level phonotactic generalizations about feature groupings and segment sequences, to higher-level relational, morphological generalizations about possible paradigmatic relationships. Because the sequences referred to by these generalizations can overlap to any degree, there is the possibility of conflict between distinct kinds of generalizations. The following section discusses the possible outcomes of this conflict in terms of competition between levels of selection.

4.6 Similarity biases and selection

In biological evolution, errors in the replication of a gene are thought to be random, at least with respect to the phenotype conferred by the gene. Selection on the basis of the interaction of a variant gene product with its environment influences the likelihood of reproduction of some unit containing the gene (such as a cell, a multicellular organism, or a kin group). As a consequence, the production of variants and the filter on what variants survive to reproduce are mechanistically distinct. On the other hand, within a model of language in which errors can be biased by similarity to other existing forms and patterns, variation in what is produced and what is perceived is nonrandom with regard to the “phenotype” of the system (cf. CHANCE and CHOICE in the framework of Evolutionary Phonology; Blevins 2004). In this regard, similarity-biased error acts in production as a selective filter acting on the pool of *potential* variants, influencing which variants actually emerge to become part of the exemplar set of the larger system. In perception, similarity-biased error acts as a selective filter by biasing identification and storage into categories.

In biological systems, genes exist within a Russian doll of nested units that are potentially the objects of selection, ranging from the gene itself, through the chromosome, the cell, the multicellular individual organism, the kin group and potentially beyond (Mayr 1997). Selection can potentially act at each of these levels, often mediated by distinct mechanisms and on different time scales.² For example, selection at the level of the cell strongly favors cells that are unconstrained in their growth, which promotes the development of cancer within an individual's lifetime. Selection at the level of the individual on the other hand strongly favors strong control over cell division. In concert, these two selective pressures lead both to selection for cancer within the population of cells within a single individual, and to selection against early development of cancer over a timescale of many lifetimes within the population of individuals (Merlo *et al.* 2006).

Within a single level of selection, the net selection pressure deriving from multiple independent loci of selection can often be approximated as a simple sum. For example, if a trait increases fitness in some way to a given degree, but decreases it by the same degree through an independent pathway, the net selection pressure on that trait may be near zero. In contrast, when selection pressures on a given trait operate at different levels of selection, say the cell versus the individual, these pressures can interact in a more complex way. Selection against a trait at one level can often proceed through creating a systemic change that makes selection *for* that trait at another level less efficient. One way to influence the efficiency of selection is through modifying the amount of variation present at a given level of selection; greater variance provides more opportunities for a fitter variant to be selected (e.g., Taylor *et al.* 2002).

Within the model presented here, the competition between selection for regularity at distinct levels of linguistic organization is similar to biological multilevel selection in that change at one level can influence the opportunities for change at another. Within the present model, a pattern serves as a self-reinforcing attractor by biasing variation/error toward itself. Because linguistic categories can overlap with or contain one another (as, for example, when a sound category is a component of a sound–meaning category such as a word), a change that increases the regularity of a pattern at one categorial level can decrease it at another. A decrease in regularity of a pattern (i.e., an increase in variance) therefore has two interacting effects on further change: (i) as variance increases at that level, the range of future variation increases,

² The well-known phenomenon of kin selection is a particular case of selection beyond the individual. In kin selection, selection at the level of the kin group favors the evolution of behavior detrimental to the self when it supports the greater reproductive success of a close relative.

potentiating change; (ii) as variance increases, similarity-driven selection pressure toward the mean is weakened. Both of these effects should independently potentiate a shift further away from regularity in the contents of a category through evolutionary change. This is illustrated in the next section.

4.7 Illustrating multilevel, selection-driven pattern development by simulation

In Wedel (2007) I illustrated the evolution of regular stress patterns through similarity-based positive feedback within a simulated lexicon over many cycles of production and perception. In these simulations, the only relationships encoded between lexical items were on the basis of shared segmental properties in temporal order. Segmental properties that were provided to the system included stress value, segmental category features and word-edge status. An example of a three-syllable lexical entry is

(5) [1, a, I] [-1, b] [1, c, F]

where square brackets enclose syllables. Each syllable is characterized by a stress value and one or more additional features: “1” and “-1” represent stress and stresslessness, respectively, lower-case letters represent segmental features, and “I” and “F” correspond to “word-initial” and “word-final”.


Lexical production in each round of the simulation proceeded by copying the information stored in the lexical entry into an output form with a low probability of error in the stress value, and then restoring it in the lexical entry, replacing the original. Directional change could intervene in this process through the action of two kinds of error-bias in output production, one external, and the other system-dependent. The external error bias was a constant, lexicon-independent bias favoring alternating stress, such that each word would eventually tend to exhibit alternating stress regardless of the initial state. The second kind of error consisted of a similarity-bias in which output stress values had a slight probability of deviating from the stored value toward the values of other forms, in relation to similarity and type-frequency. The simulation detected pattern trends within the lexicon by identifying every existing combination of features and stress values in the lexicon, and looking for robust generalizations. When an existing robust feature-set ~ stress-value generalization conflicted with the stored version of a word, the output based on that word had a greater than chance probability of shifting stress values to match the larger generalization. As a result, the system showed a strong tendency to create broad associations between stress values and features over many cycles of production and restorage.

Within the lexicon, there were many possible segmental features, but only two edge features (initial vs final). Many words therefore failed to share any segmental features at all, while every word had both an initial- and final-edge feature associated with the initial and final syllables, respectively. As a consequence, the most robust generalization that the lexicon could possibly evolve was one in which a given stress value was consistently associated with the initial and/or final word edge, rather than to some other segmental feature(s). When both even- and odd-syllable words were included in the lexicon, the dominant pattern was the evolution of an alternating stress pattern consistently aligned *either* to the initial *or* the final syllable.




To illustrate multilevel selection within this model, I modified the simulation architecture to include two optional suffix syllables for a subset of the words in the lexicon, identified with the abstract features [y] and [z] respectively. A portion of a sample lexicon is shown in Figure (3). The final syllable of every word contains a final-edge feature (F). The lexicon consists of 80 words. Half of the words in the lexicon do not have a related suffixed form, illustrated in (6a). The other half, as illustrated in (6b), appear in addition in a suffixed form. An [F] feature appears on the final syllable in all forms.³

(6) Example of a statically regular lexicon

(a) Stem-only paradigms

Stem

 [-1, a] [1, b, F]
 [-1, c] [1, d, F]

(b) Stem and Stem+Suffix paradigms

Stem	Stem	Suffix
 [-1, e] [1, f, F]	 ~ [1, e] [-1, f] ~ [1, e] [-1, f]	 [1, y, F] [1, z, F]
[-1, g] [1, h, F]	~ [1, g] [-1, h] ~ [1, g] [-1, h]	[1, y, F] [1, z, F]

The lexicon in (6) is “statically regular” with regard to stress, because all entries show alternating stress aligned to the final syllable. In this example, the stress pattern of every word can be written as [(+), −, +]. It is “relationally irregular” with regard to stress, because stems in bare and suffixed forms show

³ In order to focus the simulation on conflict between emergent phonological and morphological patterns, the development of stress associations at the final word-edge was encouraged by eliminating the initial-edge feature. Within over fifty independent trial simulations, the system always rapidly developed a stress pattern in stem-only paradigms in which stress was aligned to the final edge.

opposite stress patterns: the stress patterns of the stems in (b) are $[-, +]$ when unsuffixed and $[+, -]$ when suffixed.

The system retains the ability to detect robust static generalizations across all words in each cycle. In addition, the system has been equipped with the ability to identify the global stress-pattern relationship between suffixed and unsuffixed forms and discover any robust associations between this relationship and any existing combination of features, using a parallel computational mechanism to that used for the discovery of static generalizations (described in Wedel 2004, 2007). This latter ability allows the emergence of relational generalizations of varying specificity. For example, a maximally specific generalization would match the stress pattern of a particular unsuffixed form to its related suffixed form, whereas a less specific generalization could emerge if a number of different unsuffixed forms shared the same stress-pattern relationship with their suffixed forms. Although implemented in a computationally distinct way, this is conceptually parallel to the mechanism of relational generalization discovery in Albright and Hayes's Minimal Generalization Algorithm (2002).⁴ As before, the process of encoding an output corresponding to a stored form was subject to error biased toward existing patterns in the lexicon in proportion to similarity and type frequency. As in the single-level simulations in Wedel (2007), low-level noise was also included, in the form of a very small probability of context-free error in correctly reproducing the stored stress value in any syllable.

The result is a system that has two distinct levels of system-dependent generalization that can influence error: static generalizations at the level of features, and relational generalizations between related words, where the targets of relational generalizations contain the targets of static generalizations. Pattern competition within and between these two levels of generalization resulted in three common classes of patterns. In one class of patterns, alternating stress developed with a given stress value consistently associated with the final edge of all words, with no reference to word identity or morphological category. This is the type of pattern that emerges in the absence of any possible relational generalization linking related words. This pattern represents full regularity with regard to phonological categories, and full irregularity within each morphologically related pair, as the stress pattern for the stem in the unsuffixed form is opposite that found in the corresponding suffixed form, as in (6) above.

In a second common pattern, all two-syllable forms in the lexicon had the same stress value associated with the final edge, while each suffixed form had the opposite stress value at its final edge, thereby preserving the stress pattern

⁴ Previous work compared the mechanism of pattern discovery used here to several computational mechanisms including Analogical Language Modeling (Skousen 1989), and showed that they all produced qualitatively similar regular patterns (Wedel 2004).

of the stem. This represents full morphological, or relational regularity with regard to stem stress pattern, and full phonological, or static irregularity with regard to final-edge stress alignment. The lexicon in (7) below exhibits this pattern: stems maintain the same stress pattern whether suffixed or not (compare to (6) above). A third pattern occasionally arose in which the paradigm associated with one suffix showed phonological regularity, and the other showed morphological regularity (see cycle 990 in (9) below).

(7) Example of a relationally regular lexicon

(a) Stem-only paradigms

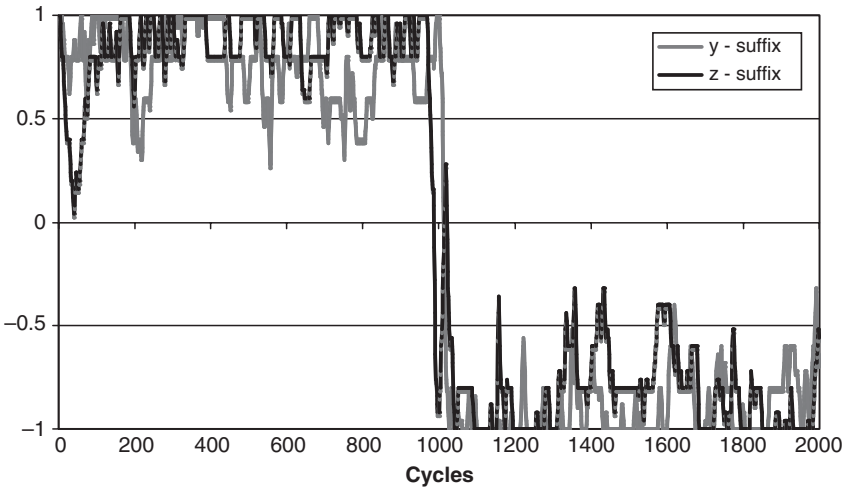
Stem	
[-1, a]	[1, b, F]
[-1, c]	[1, d, F]

(b) Stem and Stem+Suffix paradigms

Stem	Stem	Suffix
[-1, e] [1, f, F]	~ [-1, e] [1, f]	[-1, y, F]
	~ [-1, e] [1, f]	[-1, z, F]
[-1, g] [1, h, F]	~ [-1, g] [1, h]	[-1, y, F]
	~ [-1, g] [1, h]	[-1, z, F]

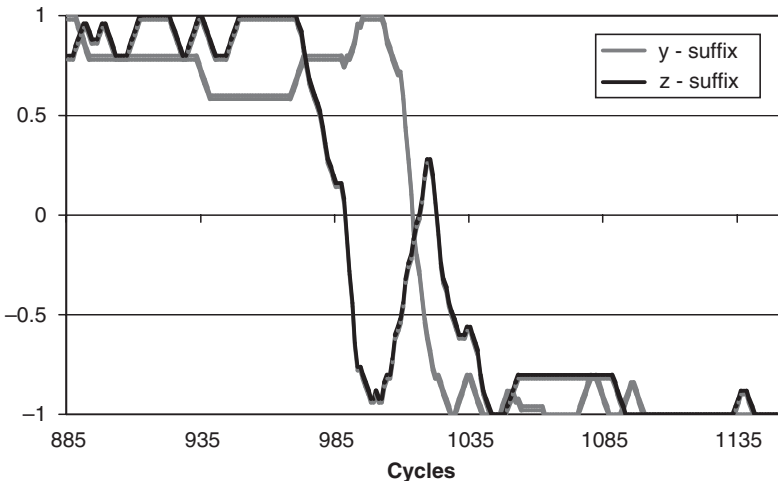
In this model, both leveling and extension occur in the same way: through errorful application of a different generalization true of some other part of the system, just as has been argued to be the case for language by, e.g., Deutscher (2001) and Hock (2003). In the event that there is only one primary relational generalization that fits all the data, then the only available mechanism for change in stress pattern lies in the low-level noise factor which provides a continual, small input of stress-pattern variants into the system. This occasionally leads to the fortuitous emergence of a different generalization, which can then spread through similarity-based error. As expected, it first spreads through the most similar subset of words within the lexicon, after which it may spread further. This is illustrated in the graph in (8) below, where a value of “1” represents full relational regularity in the stress of stems in related suffixed and unsuffixed forms, and “-1” represents full static identity in the stress patterns of all words in the lexicon with respect to the final word edge. The simulation is seeded with a lexicon exhibiting full relational regularity in both the “y” and “z” suffix paradigms, like that shown above in (4). This pattern remains stable for 1000 cycles despite the steady introduction of low-level variation in stress patterns by noise.

(8) Competition between static and relational regularities



Shortly after the z-suffix paradigm switches to a pattern in which all forms have the same stress with regard to the final word edge, the y-suffix forms are able to follow suit. This change is potentiated because the z-suffix paradigm presents a similar group of words governed by a distinct generalization which can itself be errorfully applied to members of the y-suffix paradigm. In other words, as soon as a new generalization emerges, its misapplication provides a new pathway of change. A close-up of this transition is shown in (9) below.

(9) Competition between static and relational regularities: cycle 885–1150.

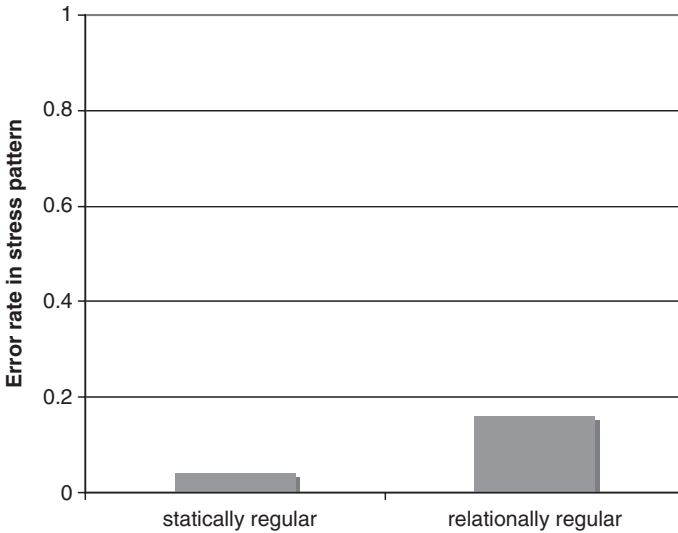


The dependence of a change between a static and relational stress pattern on the pre-existence of a target pattern in the lexicon holds in both the “extension” and “leveling” directions. In 10 simulations in which the seed lexicon was statically regular throughout (that is, where stress was aligned to the final edge of all words), it took an average of ~ 1400 cycles for a change to a relational pattern within one of the suffix paradigms to emerge. Likewise, in 10 simulations seeded with a lexicon exhibiting relational regularity in all suffixed-unsuffixed word pairs (that is, where the stem in each pair had the same stress), the average time to a change to static regularity was near 1700 cycles. In contrast, if the seed lexicon started with static regularity in one suffix paradigm, and relational regularity in the other, in 10 simulations it took on average less than 300 cycles for a further change to occur in one of the paradigms. The rate of change is greater when there are multiple existing patterns in the lexicon because each pattern represents a template for analogical extension.

The potentiating effect of multiple patterns can also be seen in the rate of error in output stress among the stem-only paradigms within the lexicon. When the stress pattern is statically regular, all stress patterns are edge-aligned across the entire lexicon. Under this condition, random noise is the only source of error in stress output in stem-only paradigms. When the stress pattern is relationally regular, however, some lexical entries have the opposite stress pattern as that in stem-only paradigms. In this case, there is an additional pattern in the lexicon to provide a pathway for variation in stress output beyond random noise. Within the simulation, this can be seen by comparing the number of stem-only paradigm outputs with a variant stress patterns in a statically regular, versus relationally regular lexicon. Figure 10 below shows the error rate in stress within stem-only paradigm outputs over 10 independent runs of 1000 cycles each in the context of either a statically, or relationally regular lexicon. When the stress pattern is consistently aligned to the final word-edge over the entire lexicon (i.e. is statically regular), the error rate in stress in stem-only paradigm outputs is .04. However, when the stress pattern is instead aligned to a stem edge within stem \sim stem+suffix paradigms (i.e., is relationally regular), the average error rate in stem-only paradigm outputs goes up to .16.

This higher error rate in the relationally regular lexicon comes about through the existence of an additional pattern in the lexicon, which provides an additional pathway to a change in stress. The resulting increased variance in stress patterns within stem-only paradigms has two related effects: (i) the dominant stress pattern of stem-only paradigms is *less* stable, and therefore more likely to change over time, and (ii), the dominant stress pattern of

- (10) Error rate in stem-only paradigm outputs given static versus relational regularity in stem \sim stem+suffix paradigms.



stem \sim stem + suffix paradigms is *more* stable, because any similarity bias promoting static regularity is weakened. This is conceptually parallel to cases in biological evolution in which selection at one level acts by modulating variance at another (e.g., Taylor *et al.* 2002).

4.8 Summary and conclusions

The statistics of error in pattern matching ensure that similar patterns will substitute for each other more often than less similar patterns in production and perception. In any model of language production and perception in which intra-categorical variants can coexist and compete within the system, positive feedback promotes the evolution of regular patterns. Under the assumption that both static and relational generalizations are manipulated during language production and perception, error between similar generalizations should produce a wide range of similarity effects at different levels, from phonotactics to morphology (Itkonen 2005).

When similarity at separate levels of organization cannot be simultaneously maximized, similarity-biased error and feedback promotes the entrenchment of one pattern at the expense of the other. The potentiation of change by similarity-biased error allows this snowball effect to proceed in opposite

directions at different levels: as variance decreases at one level, further change to solidify the spreading pattern is potentiated by feedback; at the same time, the more variance increases at the other level, the less similarity bias can work against further change. When we view similarity bias as a form of selection on the range of possible variants that enter the linguistic system over time (Wedel 2006), the interaction between overlapping levels of organization in the lexicon can be understood as a form of multilevel selection. As an illustration, I presented results from a simple simulation showing that in a system in which errorful pattern extension is a primary pathway of change, competition between a static regularity and a relational regularity resulted in the rapid stabilization of one over the other, in part by modulating variance at distinct levels of organization. Further, if a new pattern establishes itself in a subset of the lexicon, the existence of this new generalization potentiates development of a similar pattern in other, similar words. More generally, unresolved conflicts in regularity create a reservoir of instability in the system, maintaining a greater number of pathways for change and therefore a more diverse pool of variants than might otherwise exist. A motto for this model could therefore be phrased as “conflict begets variation begets extension.”

This is conceptually consistent with both the notions that analogical extension tends to result in global simplification of grammar (e.g., Halle 1962), and yet that extension is based on local generalizations (e.g., Joseph and Janda 1988, Venneman 1993). Importantly however, this model does not propose that extension serves a teleological goal of grammar simplification, but only that extension may occur more frequently when there are more grammatical patterns in competition. Multiple competing patterns provide more available pathways for error, and multiplicity itself weakens the relative strength of the behavioral attractor represented by any given pattern. Although any given analogical change may result in a relative simplification or complexification of grammar at some larger scale, any change that happens to reduce global pattern conflict also undermines the properties of the system that potentiate further analogical change. Consequently, within this model global pattern coherence is not an explicit goal of the system, but simply a relatively stable state in a continuing trajectory of change through time.

The relation between linguistic analogies and lexical categories*

*LouAnn Gerken, Rachel Wilson,
Rebecca Gómez, and Erika Nurmsoo*

5.1 Introduction

This chapter is an attempt to find points of contact between two normally distinct lines of research. One line is concerned with the psychological mechanisms that allow language learners to discover lexical categories, such as noun, verb, etc., in the linguistic input. The other is concerned with the human ability to see two domains as analogous, such that structural components of one domain align with structural components of the other domain (e.g., Gentner 1983). For example, even children are able to complete the analogies in 1a-b, below.

- 1a. chick : hen :: calf : _____
1b. dog : dogs :: wug : _____

There are at least two reasons to explore the role of our human ability to see analogies as a possible mechanism for lexical category learning. The first is that the most frequently used approach to studying lexical category learning in the laboratory employs the completion of an extended analogy, or paradigm. In such studies, experiment participants are presented with an incomplete lexical paradigm, such as the one shown in Table 5.1, below. The cells containing question marks are not presented during the initial learning phase.

Such a paradigm might reflect a variety of lexical categories. For example, the words in the top two rows might be nouns with different number markings *o* and *a*, while the bottom two rows might be verbs with different tense inflections *of* and *op*. Or, the top two rows might be case-marked

* This research was supported by NSF grant #9709774 to LAG and NIH grant #R01HD42170 to RLG and LAG. We thank David Eddington, Toben Mintz, and Royal Skousen for helpful comments.

TABLE 5.1 An example of stimuli that might be used in a paradigm completion task. Items that would fill the cells marked by “???” are withheld during training and presented with their ungrammatical counterparts during test.

blicka	snerga	pela	jica	tama	kusa
blicko	snergo	pelo	jico	tamo	???
deegof	votof	rudof	wadimof	meefof	ritof
deegop	votop	rudop	wadimop	meefop	???

feminine nouns and the bottom two rows masculine nouns. That is, without associating the paradigm with any reference field, it simply reflects a situation in which two sets of lexical items co-occur with different markers.

Once participants have been exposed to a subset of the paradigm, they are tested for their willingness to accept paradigm-conforming items that they have not yet heard, such as *kuso* and *ritop*, as well as equally new but nonconforming items, such as *kusop* and *rifo*. Participants' ability to distinguish the conforming vs nonconforming items is taken to mean that they treat the newly learned lexical items as having distinct privileges of co-occurrence, which is viewed by many researchers as the basis for lexical categories (e.g., Braine 1966; Maratsos 1982; but see Grimshaw 1981; Pinker 1982 for the alternate conception of category acquisition). Thus, both in terms of many researchers' construal of the language learners' problem, and in terms of the way in which we test for category learning in the laboratory, lexical category formation is viewed as presenting learners with an extended analogy (in the form of a morphophonological paradigm), in which generalization from the input is viewed as filling in the blank, just as in Table 5.1 above.

The second reason to explore the role of our analogy-making capacity as a possible mechanism for lexical category learning is that, as we will document more fully below, adults and infants are unable to complete paradigms like the one shown in Table 5.1 without some additional cues to category structure. We will review the types of additional cues that have been explored by researchers, including ourselves, in the next two sections. The notion that we will explore in the final section is that the psychological mechanisms by which additional cues promote category learning is one of analogy-making. To foreshadow, the argument that we will entertain is that learners require a sufficient amount of similarity among items in a paradigm to complete the paradigm and to infer lexical categories. Put simply, we suggest that learners are more likely to detect the analogy in examples like (2b) than they are in (2a), and to use that analogical structure to group the stems in (2b) into the same lexical category.

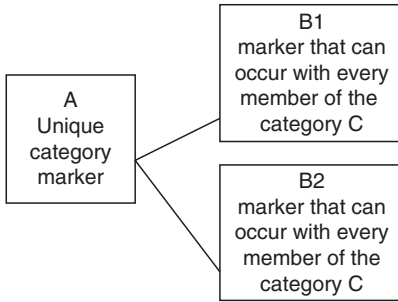
- 2a. blicka : blicko :: kusa : _____
2b. tivorblicka : tivorblicko :: tivorkusa : _____

5.2 Previous explorations into paradigm completion

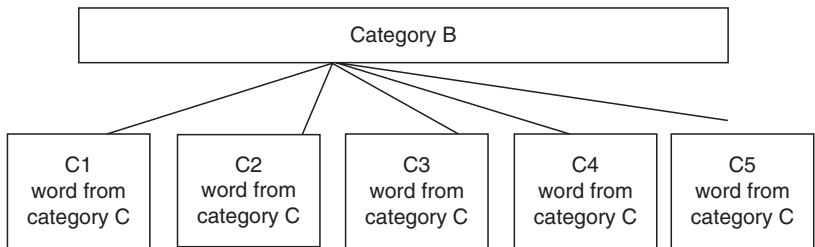
The earliest studies asking whether adults are able to complete four-part morphological analogies were done several decades ago (Braine 1966; Smith 1966). Smith (1966) asked whether adults could show evidence of learning categories by presenting them with a paradigm-completion task of the sort outlined in conjunction with Table 5.1. Participants were familiarized for one minute with 12 bigrams in which the letters came from four classes that Smith called M, N, P, and Q. Bigrams either had the form MN or PQ. Some of the possible MN and PQ pairings were withheld. At test, participants generated a number of incorrect strings of the MQ or PN type, suggesting that they had not kept the categories separate. That is, participants learned that M and P come first and Q and N second, but not that there are co-occurrence restrictions.

Braine (1987) dubbed the errorful performance of Smith's participants the "MN/PQ problem" and hypothesized that simple co-occurrence information alone is insufficient for humans to form categories. In a second class of accounts of how lexical categories are acquired, he hypothesized that, if referential information was included in addition to distributional information, categories may be learnable. In one study testing this hypothesis, Braine (1987) presented participants with an MN/PQ type language, now with MN and PQ each a phrase comprising two auditory nonsense words. Each phrase was presented with a picture. Half of the N words were accompanied by pictures of women and half of the Q words by pictures of men. The other half of the pictures depicted inanimate objects with no apparent referential regularity. Additionally, the M and P words corresponded to cardinality in the pictures. Thus, there were M words for 'one' and 'two', and P words for 'one' and 'two'. As in the work by Smith, some of the possible MN and PQ pairings were withheld.

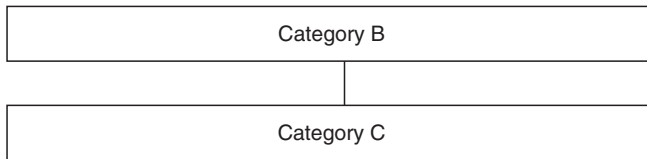
Participants made grammaticality judgments of the phrases, and unlike the participants in Smith's study, they were correctly able to distinguish unrepresented paradigm-conforming from nonconforming cases. Generalization was even correct when the phrase corresponded to a picture of an inanimate object, suggesting that participants had formed categories of M, N, P, and Q words and did not need reference to access the categories, once they had been formed. Braine speculated that the mechanism learners used to solve the



Stage 1, Unique marker A associated with items B1-Bn



Stage 2, items B1-Bn are treated as a category due to their association with unique marker A



Stage 3, Items C1-Cn are treated as a category due to their association with Category B

FIGURE 5.1 Schematic of Braine's (1987) conception of category formation from morphophonological paradigms. The presence of the A element is critical using the morphophonological markers in B with the lexical category C.

MN/PQ problem was to note first that N and Q words behave as categories, based in this case on referential information. Learners then noted that the referentially based categories co-occur with morphophonological markers (in this case, number words differentially marked for gender), and ultimately used the markers themselves as the basis of categorization. Note that Braine's account holds that learners first discover the existence of categories based on a unique cue to each category (e.g., referring to masculine vs feminine pictures; see Fig. 5.1 for a schematic). Braine also suggested that reference was not necessary for learners to hypothesize the existence of categories and that additional morphological or phonological cues might work as well.

Other researchers began to test Braine's claim that reference might not be necessary for adults to be able to complete a morphophonological paradigm, but rather that phonological or morphological information could be used to first establish the existence of categories. Until recently, the results of these investigations have been equivocal. In one study by Brooks and colleagues, participants were familiarized with one of two artificial languages (Brooks, Braine, Catalano, Brody, and Sudhalter 1993). The languages each comprised 30 words for objects (half in each of two categories), two sets of three locative suffixes (each set went with one of the object categories) and one agent (the subject of all the sentences and not relevant here). Note that the locative suffixes are the marker elements. The difference between the languages was that, in the experimental language, 60 percent (18 of 30) of the object words contained the syllables *oik* or *oo*, depending on their category. In the control language, no common phonetic information occurred on the object words. The question was whether *oik* and *oo* would function like the referential information in Braine's earlier studies. An experimenter acted out each phrase for the participant with props; however, the props and actions did not contain category information. Two results are of interest for the current discussion. First, participants trained on the experimental language recalled significantly more items than those trained on the control language. Second, participants trained on the experimental language generalized more to withheld items, as shown by more generation of these items in the recall task. However, generalization to items that did not contain *oik* or *oo* was equivocal. Therefore conclusions about whether salient nonreferential information can trigger category learning must remain speculative based on these investigations.

Frigo and McDonald (1998) continued to explore correlated cues to category learning. In their experiments, they told the participants that they would be learning two kinds of greetings: two of them were to be used in the evening and two during the day. The participants were also told that there were two groups of people and that one set of greetings was used in front of the ten names of people in one group and the other set of greetings before the ten names of people in the other. Thus, the two greetings for each group were equivalent to the two gendered cardinalities in the study by Braine (1987). The task for the participants was to correctly categorize which greetings went with which names. A subset of the names was distinguished with phonological markers, whereby 60 percent (6 of 10) of the names of members of one group shared a sequence of sounds, making this study similar to Brooks *et al.* (1993). Like the earlier study, Frigo and McDonald found in two experiments that participants were correctly able to generalize to unrepresented items when those

greeting-name combinations contained the phonological marker. However, when the unrepresented greeting-name combinations did not contain the phonological marker, participants performed no better than chance at associating the proper greeting with the person. In a third experiment, Frigo and McDonald placed phonological markers at only the beginnings of names, only the ends, or at both the beginning and end. Participants were not able to generalize to unstudied, unmarked forms unless the markers were salient (at least a syllable in length) and redundant (appeared at the beginning and the end of the word). Even the latter finding was weak, because it was not significant in the analysis by items. What is most striking about these results is that participants knew in advance how many categories there were. They were told that there were two kinds of greetings for two groups of people, and given distributional information that correlated with that number of categories. One might think it would have been easy for participants to form the categories and generalize to new cases, but it was not.

Kempe and Brooks (2001) examined a natural language, Russian, that has gender-based categorization of nouns. They noted in the CHILDES database (MacWhinney 2000) that language directed at children contains diminutive suffixes on 35–40 percent of nouns. This percentage differs from adult-directed speech where it was estimated that 2.7 percent of nouns are diminutives. Kempe and Brooks hypothesized that diminutive suffixes may serve to mark categories in the same way that the suffixes *oik* and *oo* did in the earlier study by Brooks *et al.* (1993). To test this hypothesis, they presented adult participants with two-word phrases of Russian, consisting of a color word (either masculine or feminine) and a noun. Color words served as markers. Half of the participants were familiarized with diminutivized nouns, and half were trained on the same nouns without suffixes. Phrases were presented with pictures, but the pictures added no category information. The group trained on diminutivized nouns outperformed the other group in a recall task. However, as in Brooks *et al.* (1993), generalization to phrases not containing diminutives was weak.

There are at least two explanations of the equivocal results found by Brooks *et al.* (1993), Frigo and McDonald (1998), and Kempe and Brooks (2001). The first is that human adults find it extremely difficult to form categories that are not at least in part referentially cued. However, a second explanation arises when one considers that participants in these three studies were presented with a referential field for each familiarization phrase. Importantly, the referential field did not contain any category information. Therefore, participants may have focused more fully on learning the referents for words in the familiarization phrases than on the structure of the language. The second explanation maintains the viability of Braine's original view that any add-

TABLE 5.2 Critical stimuli from Mintz (2002). The item that would fill the cell marked “???” was presented at test along with a new ungrammatical foil. The empty cell in the lower right was not presented at all during the study.

bool nex jiv	bool kwob jiv	bool zich jiv	bool pren jiv
sook nex runk	sook kwob runk	sook zich runk	???
zim nex noof	zim kwob noof	zim zich noof	zim pren noof
poz nex fen	poz kwob fen	poz zich fen	poz pren fen
choon pux wug	choon yult wug	choon plif wug	

itional pair of cues to the existence of categories might allow paradigm completion. However, tests of the hypothesis must take into account the possibility that the presence of referential information that is not relevant to the categories may disrupt category formation.

Mintz (2002) used a version of paradigm completion with adults and showed a much clearer ability to complete the paradigm based on linguistic (non-referential) cues alone. As in the examples we have been considering, Mintz employed two categories of words – those in rows 1 to 4 of Table 5.2 and those in row 5. As in the examples that we have considered, Mintz asked adults about their acceptance of the cell withheld during the initial exposure (*sook pren runk*) and compared that to their acceptance of another novel sequence (*choon pren wug*). Specifically, he asked participants to say whether or not test items were grammatical and also how confident they were in their ratings. He found that, when grammaticality ratings (1 or -1) were multiplied by confidence ratings (1–7), adults had significantly higher ratings for the paradigm-conforming than the nonconforming stimuli.

One construal of Mintz’s findings is that adults used the two simultaneously present cues to categories (e.g., *bool-jiv* vs *choon-wug*) to divide the medial words into two categories. Therefore, the study demonstrates that paradigm completion can be achieved in the absence of referential information. However, Mintz’s paradigm is not typical of the other studies using the paradigm-completion approach. Rather than presenting participants with two categories, each with two sets of markers, and testing them on unrepresented cells from each category (as in Table 5.1), Mintz presented one category with four fully correlated pairs of markers (rows 1–4 of Table 5.2) and one category with one fully correlated pair of markers (row 5 of Table 5.2). Further, Mintz only tested participants on the unrepresented cell of one of the two categories. It is difficult to say how this particular approach to paradigm completion might have yielded different results than the more traditional approach shown in Table 5.1 and used by most other researchers. Nevertheless, it would be helpful to know if learners can show clear evidence of paradigm

completion when exposed to the more traditional paradigm-completion task without referential cues. A set of studies from our own laboratory using the traditional task is presented in the next section.

In addition to its unusual use of the paradigm-completion task, Mintz's study does not appear to support Braine's construal of the role of additional cues to category membership. Recall that Braine hypothesized that the presence of a pair of cues to categories (e.g., masculine and feminine referents, the syllables *oo* and *ee*) alerted participants to the existence of two categories, which they then more fully filled in from the paradigm. However, there does not appear to be such a pair of cues in Mintz's paradigm. Rather, as noted above, four fully correlated pairs of syllables mark one category and one pair marks the other.¹ Perhaps participants could have noted a category that always began with *choon* and used that to initially form a *choon* category and an "other" category. However, Mintz's success at finding paradigm completion in a situation at least superficially different from the one characterized by Braine as learnable suggests that the psychological mechanism underlying paradigm completion may be different from the one envisioned by Braine. We will explore analogy-making as the alternate mechanism in the last section of this chapter.

5.3 Work on paradigm completion from our laboratory²

This section presents four studies that demonstrate adults' and infants' ability to complete paradigms based on linguistic cues alone, as long as just a subset of the input is marked with an additional pair of cues to category membership.

Adult Experiment 1

The stimuli consisted of 12 words of Russian, six masculine and six feminine, each with two different case endings (see Table 5.3, below). The case endings in this experiment were *oj* and *u* on feminine nouns and *ya* and *yem* on masculine nouns. Three of the feminine nouns shared a common derivational suffix (-k) and three of the masculine nouns shared a common derivational suffix (-tel). These derivational suffixes constituted partially correlated phonological information. Note that phonological information was presented on 50 percent of the words, the same percentage used by Braine (1987) for referential cues, and the same as or lower than the percentages in the other studies reviewed above. The 12

¹ Mintz (2002) referred to his stimuli as being composed of three words. They were read with the intonation of the sentence *I see you*.

² Experiments 1–3 were part of the Ph.D. dissertation of Rachel Wilson (2002).

TABLE 5.3 Stimuli used in Experiment 1. The items that would fill the cell marked “???” were presented at test along with new ungrammatical foils.

Feminine Words					
polkoj	rubashkoj	ruchkoj	???	knigoj	korovoj
polku	rubashku	ruchku	vannu	knigu	korovu
Masculine Words					
uchitel'ya	stroitel'ya	zhitel'ya	???	korn'ya	pisar'ya
uchitel'yem	stroitel'yem	zhitel'yem	tramvayem	korn'yem	pisar'yem

words each with two case inflections yielded 24 possible stimuli. However, following the typical paradigm-completion design, one feminine and one masculine item were withheld during familiarization to be presented at test, yielding 22 stimulus items presented during familiarization.

Familiarization stimuli consisted of 22 Russian words in four different random orders and presented across four blocks of trials, for a total of 88 stimuli. The stimuli were recorded by a fluent, non-native speaker of Russian (RW). Two seconds of silence was inserted between adjacent items. The familiarization session lasted a total of approximately four minutes. The critical test items were the unprecedented paradigm-conforming items *vannoj* and *tramvaya* and the equally new but unconforming items *vannya* and *tramvayoj*. An ANOVA on the number of “grammatical” responses to these two types of items was significant ($F(1, 15) = 25.90, p < .001$; see Fig. 5.2).

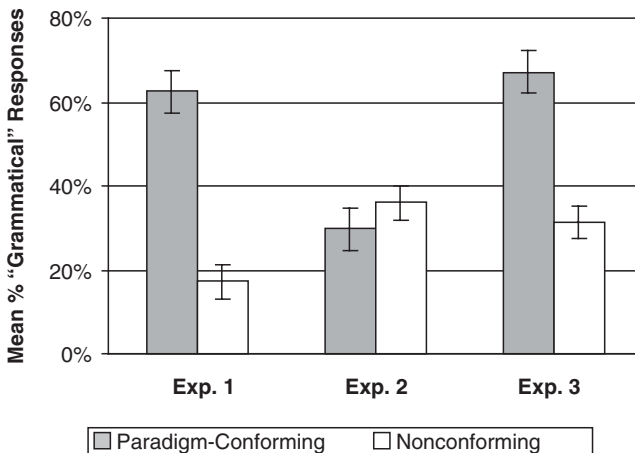


FIGURE 5.2 Mean “grammatical” responses and standard errors in Exps. 1–3. Lexical categories were marked with partially correlated cues in Exps. 1 and 3, but only case-marking cues in Exp. 2.

The data from Exp. 1 suggest that adults are able to successfully complete a linguistic paradigm based on morphological and phonological cues alone. Therefore, they are consistent with the data from Mintz (2002), but demonstrate paradigm completion using a more typical paradigm than the one used by Mintz. This study also raises two questions: First, as the literature review in Section 5.2 revealed, finding evidence of category learning by adults has been extremely difficult, and there has been no evidence at all that adults can learn when categories are marked with a single morphosyntactic cue. Indeed, our research team has failed in every one of our varied attempts to find category learning based on a single cue. Therefore, in order to further support the view that at least a subset of the words in a paradigm, such as the one shown in Table 5.3, must have two simultaneously available cues to lexical categories, we examined category learning in the presence of a single cue in Exp. 2.

The second question raised by Exp. 1 concerns the possibility of an unintended phonological cue to categories. All of the consonants that precede the inflectional ending in the feminine set are nonpalatalized. In contrast, all of the consonants preceding the inflectional ending in the masculine set are palatalized. Palatalization of the last consonant affects the shape of the inflectional ending in Russian, thereby providing additional correlational information for the learner. Put another way, because the masculine words ended in palatalized consonants, the feminine case endings that were added to these words in the ungrammatical test items sounded like *yu* and *yoj*. But in familiarization and in the grammatical test items, these endings sounded like *u* and *oj*. This additional cue might have helped participants discriminate grammatical from ungrammatical test items. This potential confound was eliminated in Exp. 3.

Adult Experiment 2

Exp. 2 presented adults with the same types of Russian words used in Exp. 1. However, in Exp. 2, the only cues to gender categories were morphosyntactic case endings. Based on the existing literature, and unpublished studies from our laboratory (Gerken, Gómez, and Nurmsoo 1999) we predicted no category learning. Because the materials for Exp. 1 had phonological cues that partially correlated with the morphosyntactic markers to gender categories, the same stimuli could not be used in Exp. 2. Therefore, a different set of 12 Russian words, each with two case endings, was used (see Table 5.4, below). These words did not provide a phonological cue to gender categories. In all other respects, Exp. 1 was identical to Exp. 2.

TABLE 5.4 Stimuli used in Experiment 2.

Feminine Words					
malinoj	rubashkoj	lapoj	vannoj	knigoj	korovoj
malinu	rubashku	lapu	vannu	knigu	???
Masculine Words					
dekana	mal'chika	vora	shkafa	plakata	brata
dekanom	mal'chikom	vorom	shkafom	plakatom	???

An ANOVA on the “grammatical” responses to the unrepresented paradigm-conforming vs nonconforming items showed no difference ($F < 1$; see Fig. 5.2). Additionally, a 2 experiment (Exp. 1 vs Exp. 2) \times 2 grammaticality ANOVA showed a significant interaction between experiment and grammaticality ($F(1, 30) = 19.77, p < 0.001$), such that only Exp. 1 participants engaged in successful paradigm completion.

Adult Experiment 3

Recall from the discussion of Exp. 1 that the goal of Exp. 3 was to rule out the possible confound of palatalization. The familiarization stimuli and paradigm-conforming test items were identical to those used in Exp. 1. However, paradigm-nonconforming test items were replaced, such that the test items that had been palatalized (e.g., *vannya* and *tramvayoj*) were now unpalatalized. An ANOVA on “grammatical” responses to paradigm-conforming and nonconforming test items was also significant ($F(1, 15) = 17.75, p < .001$; see Fig. 5.2). The fact that the results are the same in Exps. 1 and 3 suggest that the potential confound of palatalization noted with respect to Exp. 1 is not responsible for adults’ successful paradigm completion.

The data from Exps. 1–3 reported here clearly show that adults can successfully complete a morphophonological paradigm without reference, but only when the paradigm includes an additional cue that is simultaneously present with the morphological cue on a subset of the items. A remaining question from these studies is whether infants show the same ability.

Infant Experiment

Gerken *et al.* (2005) reported on a series of experiments using the Russian gender paradigm described in Adult Exps. 1–3 above. The final experiment of the published set is the most interesting for the current purposes. In it,

17-month-old infants were exposed to six masculine and six feminine Russian nouns, each with the same two case endings used in the adult studies (*oj, u, ya, yem*). Two groups of infants were tested. One group was familiarized for two minutes with words in which a subset (three feminine and three masculine) included the additional phonological cue to gender (*-k* for feminine words and *-tel*) for masculine words (see Table 5.3). This group received the same stimuli used in Adult Exp. 3, with the palatalization cue removed. The other group of 17-month-olds was familiarized with words in which none of the items had the additional phonological marker for gender. Out of the 24 possible training words (six feminine, six masculine, each with two different case endings), four words were withheld and used as the grammatical test items. Four ungrammatical words were created by putting the incorrect case ending on the four stems used for the grammatical items. Both groups of infants were tested on the paradigm-conforming and nonconforming items. Infants were familiarized and tested using the Headturn Preference Procedure (Kemler Nelson *et al.* 1995). The results mirror the findings in the adult studies – infants who were familiarized with the words with partially correlated cues to gender discriminated grammatical vs ungrammatical items at test. In contrast, infants familiarized with words exhibiting only the case-ending cue to gender failed to discriminate the test items. Further, the drop-out rate among the latter group was significantly higher, suggesting that even during familiarization, they found it difficult to discern any pattern in the stimuli.

Although 12-month-olds were not successful at learning the Russian gender paradigm that 17-month-olds so readily learned, Gómez and Lakusta (2004) reported on an apparent precursor to the category learning by 12-month-olds. They asked if these infants could learn the relationship between specific a- and b-words and features defining X- and Y-categories. During training infants heard one of two training languages. One language consisted of aX and bY pairings, the other of aY and bX pairs. Xs were two-syllable words and Ys were one syllable so that infants could use syllable number as a feature for distinguishing X- and Y-categories (e.g. *erd-kicey, alt-jic*). At test, infants trained on aX and bY pairings had to discriminate these from aY and bX pairs. However, in order to assess generalization, all X- and Y-words were novel. The infants successfully discriminated the legal from illegal pairs, suggesting that they had learned the relationships between the a- and b-elements and the abstract feature characterizing X- and Y-words (syllable number).

The difference between 12-month-olds and 17-month-olds appears to be that the latter group is able to create a cross-utterance association among “a” items and among “b” items. In terms of the Russian gender paradigm,

17-month-olds must have formed an association between *oj* and *u* and between *ya* and *yem*. This association is what allows them, upon hearing *pisarya* to know that *pisaryem* is likely. Twelve-month-olds do not yet appear capable of this cross-utterance association.

5.4 A role for analogy in lexical category learning

Let us begin with a quick summary of three main points made in the preceding sections: First, many of the studies in the literature that have examined lexical category learning by adults, children, and infants, have employed a paradigm-completion task. Many researchers, including us, view the completion of morphophonological paradigms as an important component of lexical category learning in natural language. That is, the paradigm-completion task reflects, for many researchers, not just a useful experimental task, but a task that is close to the one faced by real language learners.

Second, a growing number of studies suggest that adults, children, and infants are unable to successfully complete morphophonological paradigms of the sort illustrated in Table 5.1 unless there are additional cues to category membership presented on at least a subset of the items. In the study by Braine (1987), the additional cue came from referential categories (masculine and feminine people) that were associated with a subset of the lexical items. In Mintz (2002), the additional cue appears to be the presence of correlated syllables flanking the syllable that is crucial at test. In the Russian gender studies conducted in our laboratory, the additional cue is a phonological marker presented on a subset of the to-be-categorized word stems.

Third, Braine (1987) proposed that the function of the additional cue is to provide initial evidence of two (or more) categories, which learners then come to associate with morphophonological markers, which ultimately become the basis of the categories. Braine's own work is consistent with this view, as are the data from our Russian gender studies. In the latter studies, the phonological markers *-k* and *-tel* could, on Braine's account, serve to inform learners that there are two categories, which they then more fully discover via the case endings, finally extending the categories to items not containing *-k* or *-tel*. However, as we noted in discussing the Mintz (2002) research, that study does not lend itself as easily to Braine's account of the basis of paradigm completion or lexical category learning.

In the remainder of this section, we explore the possibility that the role of additional cues to lexical categories is in analogy-making. The presence of two simultaneous cues to category membership might be related to at least two approaches to analogy-making. One approach concerns the observation by

several analogy researchers, who suggest that the “goodness” of analogy is determined by the number of points at which one domain can be aligned with another (“structural alignment”, e.g., Gentner 1983; Gentner and Markman 1997; Holyoak and Thagard, 1997). In the Introduction, we used examples (2a) and (2b), repeated below as (3a) and (3b), to foreshadow this point. We suggested that we might feel more confident in our response to the analogy in (3b) than (3a), perhaps because (3b) has more alignable elements and therefore constitutes a better analogy than (3a). On the analogical alignment view, categories might arise as clusters of items that participate in the same high-quality structural analogies.³

3a. *blicka* : *blicko* :: *kusa* : _____

3b. *tivorblicka* : *tivorblicko* :: *tivorkusa* : _____

Another, more specifically linguistic approach can be found in the work of Skousen (1989; this volume) and other researchers using his models (e.g., Eddington 2002; Elzinga 2006). Within the model, a database is searched for items similar to a given form based on a set of potentially shared features. Items sharing particular subsets of features are grouped into sets called “supracontexts”. A subset of the supracontexts that meet a test for homogeneity (Skousen 1989) provides potential analogical models. One of the properties determining whether a supracontext will serve as the basis of analogy-making is “proximity”, in which items from the database that share more features with the given form will appear in more supracontexts and will therefore have a greater chance of being used as an analogical model (Skousen 1989). In (3a-b) above, *tivorblick* and *trivorkus* should appear in more supracontexts than *blick* and *kus*.

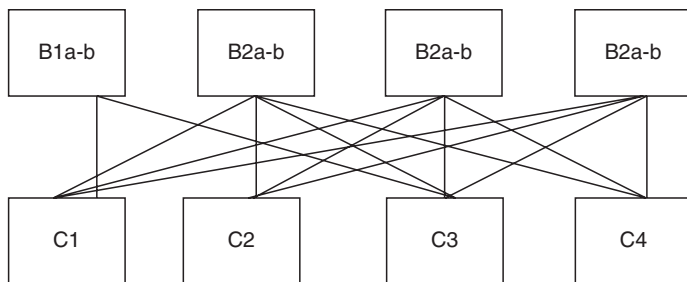
Applying this model of analogy-making to the question of lexical category formation, a lexical category is a set of lexical items that shares a large number of supracontexts (e.g., phonological, morphological, syntactic, semantic). Discovering lexical categories in paradigms such as the ones we have been discussing is facilitated when some cells in the paradigm share a large number of supracontexts, such as Russian feminine nouns that end in *-k* plus *oj*. Essentially, the strength of proximity effects can be determined by looking

³ An alternative to the structural alignment view of analogy-making comes from work on similarity-based or Bayesian category induction (e.g., Hahn and Chater 1998; Tenenbaum and Griffiths 2001). Here, the greater number of overlapping syllables (and possible morphemes) in 3b than 3a make it more likely that ‘tivorblick’ and ‘tivorsnick’ belong to the same lexical category and therefore are able to participate in the same morphological paradigm. Thus, on this view, category membership is determined separately and contributes to the evaluation of similarity.

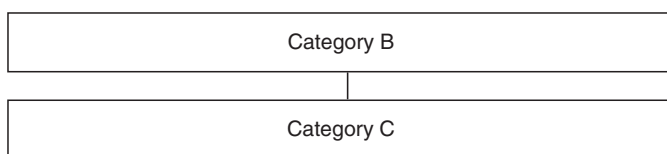
for the degree of similarity across rows of paradigms like the one shown in Tables 5.1–4.

Although proximity might explain why paradigms with two or more simultaneous cues to category membership are more likely to allow paradigm completion, this principle by itself does not appear to explain how items sharing fewer supracontexts (e.g., feminine nouns without the *-k* marker but with *oj*) come to be included in the category. “Gang effects” might be invoked here, which is a property by which, if a group of similar examples behaves alike, the probability of selecting one of these examples as an analogical model is increased (Skousen 1989). An example of similar behavior from the Russian gender-category-learning studies might be that words ending in *oj* also end in *u*, and words ending in *ya* also end in *yem*. Thus, morphological paradigms are essentially “gangs”. The strength of gang effects can be determined by looking down the column of a paradigm like the ones shown in Tables 5.1–4; the more rows there are in the column, the greater the gang effects.

So far, our attempt to view the Russian gender data in terms of analogical modeling does not seem conceptually very different from the view proposed by Braine (1987; see Fig. 5.1). The important observation we are making here, though, is that both proximity and gang effects contribute to the process by which a learner is able to demonstrate category learning by completing a paradigm. That is, unlike in Braine’s proposal, there is no need for category discovery via a unique marker (such as the *-k* ending on a subset of Russian feminine nouns). Rather, this marker can contribute to proximity effects, but its uniqueness to the category may not be necessary. This observation raises the interesting possibility that paradigms with relatively strong proximity and gang effects properties are discoverable even when no unique marker is present. Perhaps what we are seeing in the paradigms are the kinds of effects discussed by Mintz (2002, see Fig. 5.3). Note in Table 5.2 above, which shows the paradigm used by Mintz, that proximity effects are very strong: each of the four items in a row contains exactly three syllables and, importantly, begins and ends in the same syllables. In contrast, the words used on the Russian gender paradigm shown in Table 5.3 range from two to four syllables, and syllables in a row share either one or two morphophonological markers (e.g., *-oj* and *-k* on some items). Mintz’s stimuli also show strong gang effects: each column of the tested category contains four rows, in contrast with the two rows used in the Russian gender paradigm work and most other paradigm-completion studies. The account given here suggests that either reducing the Mintz’ frames (rows) to a single marker element, or reducing the number of frames, should yield less successful category learning. This account also suggests that learners might be able to engage in successful paradigm



Stage 1, Frequent frames (e.g., *bool-jlv*) are associated with medial elements (e.g., *pren*)



Stage 2, Items B1-Bn are treated as a category due to their association with items C1-Cn, and items C1-Cn are treated as a category due to their association with items B1-Bn.

FIGURE 5.3 Schematic of category formation in the experiment by Mintz (2002).

completion in a Russian gender paradigm with only case-marking cues (e.g., Table 5.4) if they heard each noun with four different case endings. Such a finding would provide strong evidence against Braine's view of category formation. We are currently undertaking the relevant experiment in our laboratories. Interestingly, we might use the notion of proximity and gang effects properties to account for the difference we observed between 17- and 12-month-olds. Recall that, while the older group was able to complete the Russian gender paradigm, the younger group was not. However, the younger group was able to associate a marker element with number of syllables in the adjacent word. Perhaps the younger infants were only able to employ the proximity principle in their creation of protocategories, while older infants, as we suggested above, were able to use both principles under discussion.

Let us end by acknowledging two negative points about the proposal we have sketched here. One point applies specifically to viewing lexical category learning as analogy-making. Skousen's model has been applied to a number of synchronic and diachronic linguistic problems with good success. It appears

to offer an interesting way of thinking about lexical category formation that we hope researchers will pursue. However, an analogical approach to category formation would need to utilize a wide range of features in computing supracontexts, including reference (as seen in the Braine's (1987) study with masculine and feminine referents) and words in phrases (Mintz 2006). The latter point suggests that the database from which supracontexts are computed must be something other than the lexicon. In short, the computational problem suggested by our proposal may simply be intractable (see Skousen in this volume).

The second potential negative that we should acknowledge applies to all approaches to lexical category formation that are driven by distributional cues (phonological, morphological, and sentence-structural contexts). The categories formed by such an approach cannot be easily linked to labels such as “noun”, “verb”, etc. Rather, they are simply groups of words that share similar properties. Insofar as linguistic theories require learners to have innate category knowledge, the category-formation mechanism that we are exploring is problematic (see Gerken *et al.* 2005 for further discussion). Conversely, insofar as a distributional model of category learning can be shown to be successful for accounting for human learning, we may need to abandon notions of innate categories in favor of some form of guided category learning. We hope that viewing category learning within an analogy-making framework can contribute to this important debate.

The role of analogy for compound words

Andrea Krott

The aim of many linguistic investigations is to discover productive patterns of a language. If a pattern is very regular, it can be described by means of rules. For example, the rule for the English past tense accounts for a new form such as *wug* + *ed* > *wugged*. As in this example, novel formations are often based on patterns that are very regular and therefore can easily be described by rules. But novel formations can also be coined without the existence of a regular pattern. For instance, *ambisextrous* or *chocoholic* are based on the single exemplar *ambidextrous* and the small set of similar exemplars *alcoholic* and *workaholic*. Such seemingly accidental formations are creative and might appear exceptional. The structure or process used to explain them is analogy. Analogy is therefore sometimes viewed as an exceptional and rare process that stands in contrast to the productive formation of novel word-forms using rules (e.g., Marcus *et al.* 1995; Pinker and Ullman 2002). In contrast, some scholars view rules as extreme cases of analogy. In other words, a novel word that appears to be formed using a rule is assumed to be formed in analogy to many exemplars (e.g., Bybee 1995; see also connectionist approaches such as McClelland and Patterson 2002). While this debate predominantly concentrates on formations such as the English past tense that seemingly can be explained by both approaches, this chapter will present a type of word formation that can only be captured by analogical mechanisms, namely noun-noun compounds such as *landlady*, *airport*, or *boy scout*. I will show how analogy can systematically govern a whole category of words across different languages and how the same analogical basis can play a role in different domains of language processing, from language acquisition to visual word processing. From the presented studies it will become clear that analogy is a very powerful tool that is not rare and exceptional but frequently used and that can explain much more than accidental coinages.

In the literature, different types of noun-noun combinations have been distinguished. For instance, French has a large number of noun-preposition-noun combinations such as *sac à main* ‘bag at hand’ meaning a handbag or *chef de police* ‘chief of police’ meaning police chief as well as a small number of noun-noun combinations such as *timbre-poste* > lit. stamp-post + office, ‘a postage stamp’. There has been a discussion as to whether French has nominal compounding at all because pure noun-noun combinations are rather rare. Robinson (1979), for example, classifies only some noun-preposition-noun combinations as compounds. In English, some scholars distinguish between noun-noun phrases and noun-noun compounds (e.g., Bloomfield 1933). Stress has been viewed as the distinguishing feature between the two, with noun-noun compounds having compound stress, which is defined as primary stress on the modifier as in *bookshelf*, and noun-noun phrases having phrasal stress as in *apple pie* (e.g., Bloomfield 1933; Giegerich 2004; Lees 1960). Due to this difference it has been argued that noun-noun phrases belong to the syntax of a language, while noun-noun compounds are part of the lexicon (e.g., Giegerich 2004). However, this distinction cannot be sustained, especially because stress in noun-noun constructs is highly variable (Bauer 1998; Di Sciullo and Williams 1987; Giegerich 2004; Levi 1978; Plag 2006). Because there is no generally accepted definition of what constitutes a compound (see Fabb 1998) and because analogy – as will become apparent hereinafter – seems to play the same role for all noun-noun constructs, I will treat noun-noun constructs of different languages as a homogeneous class and refer to them as *compounds*.

The type of analogy that is important for compounds is based on similarities within sets of words rather than isolated single words. These sets are groups of compounds that share a constituent, either the modifier or the head. Sets of compounds that share a modifier are also referred to as modifier families, while sets of compounds that share a head are referred to as head families (see also de Jong *et al.* 2002; Krott, Schreuder, and Baayen 2001; Krott *et al.* 2002c). The following show the modifier and head family of the novel compound *chocolate bread*.

- (a) Modifier family of *chocolate*:
chocolate cookie, chocolate bar, chocolate cake, chocolate chips, chocolate icing, chocolate mouse, chocolate pudding, chocolate brownie, chocolate muffin, chocolate milk, etc.
- (b) Head family of *bread*:
banana bread, cheese bread, ginger bread, olive bread, rye bread, sandwich bread, wheat bread, etc.

In what follows, I will present evidence that modifier and head families play a central role in the processing of compounds. They provide an analogical basis for production, comprehension, interpretation, and acquisition of compounds for a variety of languages across different language families. Although compounding is one of the most productive word-formation processes across languages, studies of the role of constituent families to date have focused very much on Indo-European languages such as English, Dutch, German, and French. To show that we are indeed dealing with a more general phenomenon, I will also present evidence from Indonesian, Japanese, and Chinese.

6.1 Production of compounds

Maybe the strongest evidence that modifier families and head families function as analogical bases for compound processing comes from research into the production of novel compounds in Japanese, Dutch, and German, specifically the use of interfixes in these compounds as well as stress assignment in English compounds.

It has been shown that constituent families affect the choice of interfixes¹ in novel Dutch and German noun-noun compounds such as *-s-* in Dutch *schaap + s + kop* > *schaapskop* ‘sheep’s head’ or *-en-* in German *Schwan + en + see* > *Schwanensee* ‘swan lake’. The majority of Dutch noun-noun compounds, i.e., 69 percent of the noun-noun compounds listed in the CELEX database (Baayen, Piepenbrock, and Gullikers 1995), are similar to English compounds, being mere concatenations of two nouns such as *water + druppel* > *waterdruppel* ‘water drop’. However, the remaining compounds contain either *-s-* (20 percent) such as *visser + s + boot* > *visserboot* ‘fishing boat’, *-en-* (11 percent) such as *sigaret + en + etui* > *sigarettenetui* ‘cigarette case’, or in rare cases *-er-* such as *ei + er + dopje* > *eierdopje* ‘egg cup’.² Van den Toorn (1982a, 1982b) and Mattens (1984) have attempted to formulate a set of rules that capture the occurrence of interfixes by focusing on the phonological, morphological, and semantic make-up of the constituents. All of these rules, however, turned out to have exceptions. One of the phonological rules says, for instance, that interfixes should not occur after a modifier ending in a vowel as in *thee + bus* ‘tea box’, which is contradicted by a compound such as *pygmee + en + volk* ‘pygmy people’. On a morphological

¹ Interfixes are also referred to as linking elements, linking morphemes, connectives, or juncture suffixes.

² Note that the interfix *-en-* is occasionally spelled as *-e-*, but both variants are pronounced as schwa.

level, some suffixes of modifiers occur mostly with one interfix and occasionally with another. For instance, the abstract nominal suffix *-heid* occurs most frequently with the interfix *-s-* as in *snelheid + s + controle* > *snelheidscontrole* ‘speed control’, sometimes without any interfix as in *oudheid + kunde* > *oudheidkunde* ‘archaeology’, and occasionally with *-en-* as in *minderheid + en + beleid* > *minderhedenbeleid* ‘minority policy’. On a semantic level, modifiers that end in the suffix *-er* and that are human agents tend to occur with *-s-*, but see *leraar + en + opleiding* > *leraarenopleiding* ‘teacher training’. Due to the large set of exceptions van den Toorn (1982a, 1982b) concluded that there are no rules and that the regularities that can be observed are mere tendencies.

Compared to rules, analogy over constituent families has been proven to be a much more successful approach to Dutch interfixes (Krott, Baayen, and Schreuder 2001; Krott, Hagoort, and Baayen 2004; Krott, Schreuder, and Baayen 2002b). The usage of interfixes has been shown to be related to their occurrence in modifier families and head families, although the effect of the modifier is stronger (Krott, Schreuder, and Baayen 2001; Krott, Schreuder, and Baayen 2002b). In a cloze-task experiment, participants were asked to combine two nouns into a novel compound. Their responses with a particular interfix were well predicted by the support that the interfix received from the modifier family and to a lesser degree from the head family. For instance, the modifier *onderzoek* ‘research’ of the novel combination *onderzoek + schaal* ‘research scale’ occurs most frequently with *-s-* in existing compounds, while the head *schaal* is neutral in terms of occurrence of *-s-*. The results showed that 95 percent of the participants chose an *-s-*. Furthermore, constituent families not only predicted the choice of Dutch interfixes very accurately, they also predicted the speed with which they were selected (Krott, Schreuder, and Baayen 2002b). The higher the support for a particular interfix was, i.e., the higher the percentage of the interfix in the constituent families, the faster participants selected this interfix for a novel compound.

Other factors that can predict participants’ choices of interfixes are the suffix and rime of the modifier and the semantic class of the modifier. In Krott, Schreuder, and Baayen (2002a), participants were asked to create compounds from a nonword modifier and a real head as in *lantana + organisatie* ‘lantana organization’. The rime of the nonword modifier was chosen to predict the usage of different interfixes. Although participants reported higher uncertainty than for combinations of real words, the rime had a significant effect on participants’ responses. Similar results were obtained when participants selected interfixes for combinations of real heads and nonword modifiers that

ended in a suffix as in *illuni-teit + toename* ‘illuniteit increase’ (Krott, Baayen, and Schreuder 2001). Nevertheless, constituent families are the most powerful predictors of Dutch interfixes. When modifier families are in competition with the suffix or the rime of the modifier, then it is the modifier family that determines which interfix participants choose (Krott, Schreuder, and Baayen, 2002a). Thus, rime or suffix of modifiers are only fallen back on when there is no modifier family available that can provide an analogical basis for the selection. In contrast to rime and suffixes, an experiment testing the effect of semantic features of the modifier, i.e., concreteness and animacy, showed that semantic features contribute to interfix selection in addition to constituent families (Krott, Krebbers, Schreuder, and Baayen 2002).

Additional evidence for the analogical nature of the influence that constituent families have on interfix selection stems from computational simulation studies, using the exemplar-based models TiMBL (Daelemans *et al.* 2000) and AML (Skousen 1989). Constituent family effects in the behavioral experiments were all manifestations of type counts, i.e., they were based on percentages of family members with a particular interfix and not on the family members’ frequency of usage. In contrast to other types of models that implement analogical predictions such as connectionist networks, exemplar-based models easily and transparently accommodate the effect of type counts because predictions are based on explicitly stored earlier experiences, i.e., exemplars, and not, e.g., based on modified hidden nodes in a network. For instance, in Skousen’s analogical modeling of language (AML), a target word is compared with stored exemplars using a similarity algorithm defined over a series of user-selected features and then classified into a class, for instance, into an inflectional class. Exemplars that are most similar to the target provide the analogical basis for the target’s classification. The Tilburg Memory Based Learner (TiMBL) implements a similar mechanism. However, exemplars are not stored as wholes. During its learning phase, TiMBL integrates exemplars into a decision tree, which makes neighborhood searches more efficient than searching through a list of exemplars. One additional advantage of TiMBL is that it provides a measure of relevance for each user-defined feature, by calculating the information gain that the feature contributes to the prediction.

Simulation studies of the selections of interfixes in our experiments with TiMBL have confirmed the prime importance of modifier families. Modeling participants’ choices for the combinations in Krott, Baayen, and Schreuder (2001) and Krott, Schreuder, and Baayen (2002a) revealed that the modifier was by far the strongest predictor and that adding rime or suffix to the

predictive feature set did not improve the prediction accuracy when a modifier family was available. Modeling participants' responses in Krott, Schreuder, and Baayen (2002a) with AML led to equally high prediction accuracies as those obtained with TiMBL. Furthermore, comparing the models' choices with participants' choices revealed that the selection is equally difficult for human participants and the models, confirming that exemplar-based models are very good approximations of human behavior. In addition, the prediction accuracies of the models were superior to that of rules.

Similar to interfixes in novel Dutch compounds, constituent families also play an important role for the choice of interfixes in German compounds (Krott, Schreuder, Baayen, and Dressler 2007). Although German and Dutch are etymologically very close, German has a more complex system of interfixes. There are seven non-Latinate interfixes: -s-, -e-, -n-, -ens-, -es-, -er-. In addition, the modifier, i.e., the left constituent, sometimes changes its root vowel via umlaut in combination with an interfix as in *Huhn* + *er* + *ei* > *Hühnerei* 'chicken egg'.³ Other modifiers are reduced to their root as in *Farbe* + *Fernsehen* > *Farbfernsehen* 'color TV'. Similar to Dutch noun-noun compounds, 65 percent of all compounds in CELEX (Baayen, Piepenbrock, and Gullikers 1995) contain an interfix, the others are pure concatenations of nouns. While previous studies had explored the predictability of German interfixes by rules (Dressler *et al.* 2001; Libben *et al.* 2002), we focused on the prediction by analogy. In behavioral experiments we showed that the modifier family is not only a strong predictor for Dutch interfixes, but also for German interfixes. The smaller effect of the head family, which we had observed for Dutch interfixes, was less important and depended on the compound. We also tested whether we could simulate participants' selections using TiMBL. We compared the predictive power of the modifier family with that of features of the modifier, which had been proposed in a rule-based account (Dressler *et al.* 2001; Libben *et al.* 2002). The simulations confirmed the important role of the modifier family. However, they also showed that it was not the constituent family alone that best predicted the selection of interfixes. Adding properties of the modifier such as gender, inflectional class, and particularly its rime improved the prediction.

³ Compounds such as *Hühnerei* or *Wort* + *er* + *Buch* > *Wörterbuch* word + book > 'dictionary' might also be analyzed as *Hühner* + *Ei* or *Wörter* + *Buch* because the left constituent is identical to the plural form of *Huhn* 'chicken' or *Wort* 'word'. Semantically this might make sense for *Wörterbuch*, which refers to a book that contains lots of words, but not for *Hühnerei* because a chicken egg is an egg laid by one chicken only. Interfixes are therefore represented in this chapter as independent morphemes instead of plural suffixes of modifiers. Note that the effect of the constituent family is not affected by the choice of representation.

These properties did not enhance the prediction for Dutch interfixes. The difference between the languages is probably due to the overall greater importance of inflectional class and gender in German. One might argue that the predictive power of rime, gender, and inflectional class can be taken as evidence for rules and that these rules have an effect that is independent of the analogical effect of the constituent family. It is difficult to explain, however, why the effectiveness of rules and analogy varies for different modifiers (Krott *et al.* 2007). A more parsimonious account is therefore that the properties of the modifier are analogical in nature as well, i.e., their rule-like behavior is rather an extreme and highly consistent form of analogy.

Inspired by TiMBL and AML, we developed a computational psycholinguistic model of analogy that does not only predict the choices for interfixes in novel compounds, but also the speed with which participants choose (Krott, Schreuder, and Baayen 2002*b*; Krott *et al.* 2007). Like exemplar-based models, our model is a type-based, as opposed to a token-based, model of analogy, having symbolic representations of words. Figure 6.1 illustrates the connectivity between constituents, compounds, and interfixes for the Dutch example *schaap + oog* ‘sheep’s eye’. Activation initially flows from the lemma nodes of the constituents to the nodes of their constituent families and further to the interfixes that they contain. Activation only flows to the relevant family, i.e., the modifier SHEEP activates the compounds that contain SHEEP as a modifier. Note that the larger effect of the modifier family is modeled by a larger weight of the connections between the modifier and its family compared to the weight between the head and its family. In the case of German compounds, this weight would be zero for most right constituents. Activation flows back and forth between interfix nodes, compound nodes, and constituent nodes. This builds up activation in the interfixes until one of them reaches a predefined activation threshold. The time it takes to reach the threshold simulates the time that human participants take to respond. An analysis of the simulated response times showed the same effects of the modifier family and the predicted lack of an effect of the head family on response times that had been observed in the behavioral studies with human participants (Krott, Schreuder, and Baayen 2002*b*).

Research into Japanese compounds has shown that constituent families also play a role in a language of a very different kind. Japanese *rendaku* is the voicing of the initial obstruent of the second constituent in compounds or stem-and-affix formations. For example, the compound /ami/ + /to/ ‘net + door’ becomes /amido/ ‘screen door’ and /iro/ + /kami/ ‘color + paper’ becomes /irogami/ ‘colored paper’. However, *rendaku* is not applied in all cases. Lyman’s Law states that *rendaku* does not occur if the second

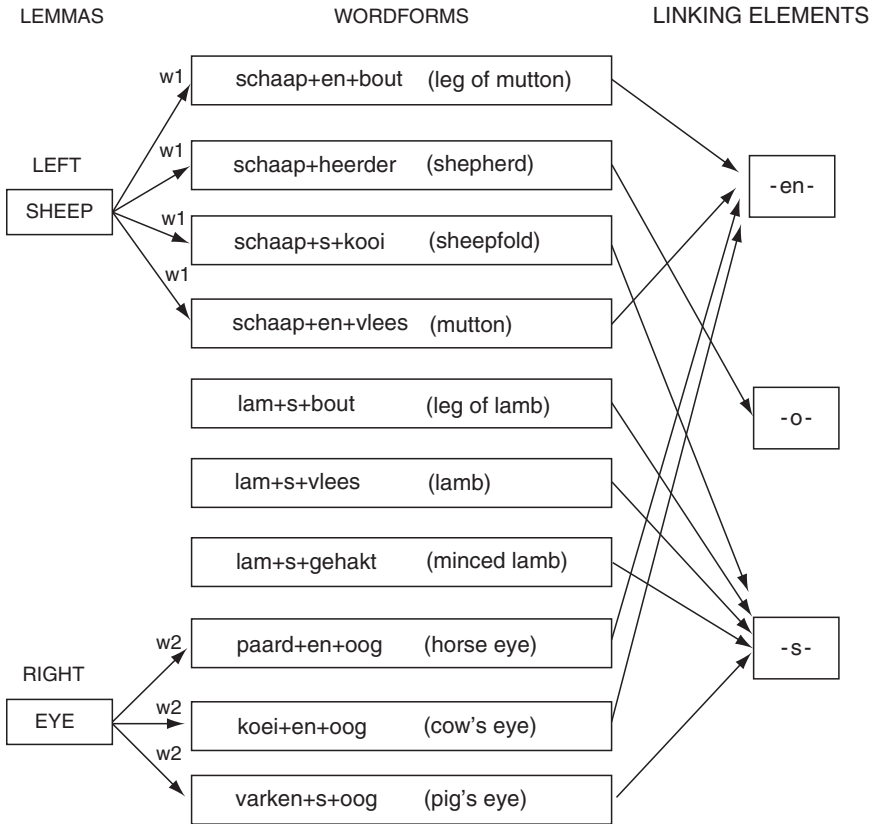


FIGURE 6.1 Connectivity of a simple Dutch compound lexicon: lemmas (left layer), word-form representations (central layer, equivalent to lexemes in Levelt, Roelofs, and Meyer's 1999 model), and interfixes (right layer), as developed by Krott, Schreuder, and Baayen (2002b).

constituent already contains a voiced obstruent (Vance 1980). Vance (1980) tested the psychological status of Lyman's law for novel compounds. He found a correlation between participants' preference for rendaku in a compound and the likelihood of rendaku in the head family of the compound. Importantly, head families had a stronger effect on participants' responses than Lyman's law. The law was only effective if the second constituent was a nonword, i.e., when there was no head family.

An effect of constituent families that is of a slightly different nature is its role for assigning stress in English compound production. As mentioned, stress in English noun-noun constructs is highly variable (Bauer 1998; Plag 2006). Plag (2006) investigated what affects this variability, by focusing on

three factors in existing and novel noun-noun combinations. The first factor was compound structure, predicting that complement-head combinations such as *opera-singer*, with *opera* as complement to *singer*, are clearly left-stressed in contrast to modifier-head compounds such as *opera glasses*. The second factor was semantics, predicting that compounds with modifier relations that express authorship such as *Groszkinsky symphony* are right-stressed, while compounds with modifier relations that express a title such as *Moonlight Symphony* are left-stressed. The third factor was analogy, more precisely the effect of the head family. For example, compounds with the head *street* are stressed on the modifier such as *Oxford Street*, *Main Street*, while compounds with the head *avenue* are stressed on the head such as *Fifth Avenue*, *Madison Avenue*. Measuring pitch differences between modifiers and heads in participants' oral compound productions, Plag found some evidence for all three factors. In terms of analogy, he found that compounds with *symphony* as head such as *Spring Symphony*, *Hoffman symphony* etc. showed a smaller pitch difference than compounds with *sonata* as heads such as *Twilight Sonata*, *Winter Sonata* etc., or *opera* such as *Surprise Opera*, *Groszkinsky opera*, etc. In addition, the analogical factor of the head family overruled the effect of the semantic relation of the compounds. For example, all compounds with the head *symphony* behaved equally with regards to stress, which is expected given the head family of *symphony*. The semantic factor predicts, however, that modifier relations that express authorship such as *Hoffman symphony* should have led to right stress and modifier relations that express a title such as *Spring Symphony* should have led to left stress. Plag's results therefore again show that constituent families can overrule other factors.

In sum, we have seen that constituent families provide powerful analogical bases for the production of compounds in a number of languages and across language families. They are highly predictive of participants' behavior, when asked to produce novel compounds. They quite accurately predict participants' decisions as well as the speed with which the participants make those decisions. Simulations with exemplar-based models provide independent support for the conclusion that constituent families are an important basis of analogical generalization in language production.

6.2 Visual processing of existing compounds

Constituent families also affect the comprehension of compounds. De Jong *et al.* (2002) studied the processing of Dutch and English compounds when

those were presented visually. They asked native speakers to decide whether or not the words appearing on a computer screen were existing words by pressing a yes or no button as fast and as accurately as possible. English compounds differ from Dutch compounds in terms of orthography. While Dutch compounds are always written as one word, English compounds can be written as one or two words, depending on the individual compound. For instance, while *heartbeat* is written as one word, *heart attack* is written as two, even though these words are very similar, sharing the modifier. De Jong *et al.* (2002) investigated whether the size of constituent families or the summed frequency of constituent family members might have an effect on how quickly speakers decide that a word is a familiar compound. Response times to Dutch compounds as well as to English compounds that were written as single words were driven by summed frequencies of modifier families. This suggests that participants' processing of compounds is sensitive to the probability that a constituent occurs as a modifier in a compound. It is likely that modifiers that occur more often are easier to recognize than modifiers that occur seldom. In case of English compounds that were written as two words, participants' reaction times reflected the size of the modifier family. Participants recognized compounds faster when the family of the modifier was large than when it was small. Head families did not appear to affect responses, which either means that they are not "active" during compound processing or that their effect is masked by an overwhelming effect of the modifier family.

Krott, Hagoort, and Baayen (2004) also investigated the processing of visually presented Dutch compounds, more specifically the support of constituent families on participants' decisions about the well-formedness of novel and existing compounds. The compounds contained interfixes that were or were not in line with the interfix bias within the modifier family. In case of existing compounds, interfixes also differed as to whether they were conventional for the particular compound as in *rat + en + vergif* > *rattenvergif* 'rat poison' or unconventional as in *rat + s + vergif* > **ratsvergif*, with the latter leading to novel compounds that are very similar to existing ones. Similar to the effects on the selection of interfixes in production (Krott, Schreuder, and Baayen 2002b), the bias of the modifier family predicted both acceptance rates and acceptance speed. More support for an interfix in the family led to higher acceptance and faster responses. Remarkably, nonconventional interfixes in existing compounds such as -en- in **kleur + en + bad* > *kleurenbad* 'color bath', were accepted as correct as often and as fast as conventional interfixes such as -en- in *dier + en + kliniek* > *dierenkliniek* 'animal hospital', as long as they had the support of the modifier family. This effect was independent of

the frequency of the compound. Thus, the modifier family determined yes responses independently of the novelty or familiarity of the compound.

Constituent families have also been shown to affect the recognition of written Chinese compounds. Similar to the paradigm used by De Jong *et al.* (2002) for English, Tsai *et al.* (2006) and Huang *et al.* (2006) investigated the effect of family size on reading Chinese compounds. Tsai *et al.* (2006) examined the effect of family size⁴ of first characters on the speed with which isolated compounds are recognized as well as its effect on eye movements when compounds are embedded in sentences. For example, the first character of 糗事 *qiǔshì*⁵ ‘dry-ration thing’, meaning embarrassing thing, occurs in a very small set of words, while the first character of 善事 *shànshì* ‘good thing’, meaning good deeds, occurs in several other words. Tsai *et al.* found that compounds with first characters that come from large families were recognized more quickly than compounds with first characters that come from small families. Eye movements revealed a higher skipping rate and shorter fixation durations for words with larger families than those with smaller ones. Both findings suggest that a large family facilitates recognition. Huang *et al.* (2006) investigated the effect of family size of both first and second characters on recognition speed, while partly simultaneously recording brain activations, i.e., ERPs (event-related encephalograms). They confirmed Tsai *et al.*’s (2006) finding that large families of first characters facilitate responses. In addition, families of first characters affected response times more strongly than families of second characters and high-frequency competing family members inhibited responses. ERPs suggested that larger families lead to increased lexical activity compared to smaller families and that a high-frequency competing family member leads to greater competition during word recognition than a low-frequency competing family member. These findings show that constituent families play a role in written compound recognition cross-linguistically. They also reveal that constituent families do not simply facilitate word recognition. A high-frequency competitor can slow recognition down.

6.3 Interpretation of compounds

Noun-noun compounds have three semantic components: a head that determines the category, a modifier that determines how the subcategory is different from other subcategories, and a relation between modifier and head. For example, an *apple pie* belongs to the superordinate category *pie*

⁴ Both Tsai *et al.* (2006) and Huang *et al.* (2006) used the term ‘neighborhood’ instead of ‘family’.

⁵ Examples are in Mandarin Romanization.

and is a pie that has apples in it, in contrast to pies that have cherries, lemons, etc. in them. Although there is in principle no limit to how nouns can be related in compounds, linguists and psycholinguists have suggested ten to twenty very common relation categories (Downing 1977; Gleitman and Gleitman 1970; Kay and Zimmer 1976; Lees 1960; Levi 1978), including the very common FOR as in *juice cup*, a cup FOR juice, HAS as in *banana muffin*, a muffin that HAS bananas in it, and MADE OF as in *carrot sticks*, sticks MADE OF carrots. To understand the meaning of a novel compound, one needs to identify its modifier and head and then to infer an appropriate semantic relation between them. Several approaches have been proposed to account for this inference process (e.g., Costello and Keane 2001; Estes 2003; Gagné and Shoben 1997; Murphy 1990; Wisniewski 1996). Most relevant for this chapter are studies by Ryder (1994), van Jaarsveld, Coolen, and Schreuder (1994), and the Competition-Among-Relations-in-Nominals (CARIN) model by Gagné and Shoben (Gagné 2001; Gagné and Shoben 1997, 2002). Ryder (1994) was the first to systematically investigate the importance of analogy for the interpretation of novel noun-noun compounds. She investigated analogical effects at various levels: specific compounds, constituent families, templates such as whole-part or container-contained, and the very general schema XY, which she defines as “an Y that has some relation to X”. She asked participants to define the meaning of novel noun-noun compounds and found that interpretations could indeed be based on all four levels of analogy. It is not clear from this research, however, what drives participants to decide which level to use.

Van Jaarsveld, Coolen, and Schreuder (1994) sought additional evidence for two of the analogy levels identified by Ryder (1994), namely the levels of specific compounds and constituent families. They constructed novel compounds with large and small sizes of constituent families. They asked participants to rate them for interpretability and then tested how fast participants recognized them as real English words in a lexical decision experiment (similar to that by de Jong *et al.* 2002). They found that compounds with larger constituent families were responded to faster than those with smaller constituent families, and the speed was independent of the compounds’ interpretability. This indicates that responses were affected by the size of constituent families, similar to the results for visual compound processing above.

Gagné and Shoben (Gagné 2001; Gagné and Shoben 1997, 2002) took a similar approach to that by van Jaarsveld, Coolen, and Schreuder (1994) in their CARIN model. They argue that the selection of a relation for a novel compound is affected by how the compound modifier has been used in

previous combinations. Therefore the availability of the relation is argued to affect the ease of interpretation. Availability is, for instance, influenced by a modifier's previous usage with a particular relation, which includes its usage in the modifier family. In contrast to van Jaarsveld *et al.*, Gagné and colleagues focused on modifier families rather than head families, and they used an experimental task that taps directly into the interpretation process. They asked participants to decide whether a visually presented novel compound made sense or not. The most likely modifier-head relation of a novel compound was either supported by a strong bias towards this relation in the modifier family or not. Participants were faster to accept the novel compound as making sense when the modifier family strongly supported its modifier-head relation than when it did not support it (Gagné and Shoben 1997). Head families affected response times only when the novel combination was ambiguous as for *student vote*, which can be a vote for students or by students (Gagné and Shoben 2002). A subsequent study showed that modifier-relation pairs can even prime sense-nonsense decisions to familiar compounds (Gagné and Spalding 2004), suggesting that modifier families are activated not only for novel compounds, but for all compounds. The findings of Gagné and colleagues for English modifier families have since been confirmed by Storms and Wisniewski (2005) for Indonesian, i.e., a language with left-headed compounds (see also a study on French compounds by Turco; as cited in Gagné and Spalding 2006). Modifier and modifier families therefore seem to play an important role in compound interpretations cross-linguistically. The role of the modifier lies in the crucial information that it provides to distinguish a particular compound from others within the same category. Part of this distinguishing information is the relation that holds between modifier and head.

6.4 Acquisition of compounds

The studies reviewed so far all dealt with the role of constituent families in compound processing in adults, i.e., in participants who have mastered the production, recognition, and interpretation of compounds. The question arises when do constituent families become effective during language development? Do children who have a limited vocabulary already make use of constituent families? In order to answer these questions, one needs to study compound processing by young children. To be able to place the emergence of the importance of constituent families into children's development, I will first give a brief description of what we know about compound acquisition. I will

then present three studies that show how constituent families are already important for 4- and 5-year-old children.

The literature on compound acquisition presents contradictory findings. On the one hand, compounds and the system of compounding appear to be learned very early. There is evidence that children start to spontaneously coin their first novel compounds such as *nose-beard* or *car-smoke* around the age of 2 (e.g., Becker 1994; Clark 1983). Two-year-olds also seem to understand already the different roles of heads and modifiers (Berman and Clark 1989; Clark 1981, 1983; Clark and Berman 1987; Clark, Gelman, and Lane 1985; Mellenius 1997). Furthermore, 3-year-olds can use compounds to refer to subcategories, suggesting that they understand the subcategorization function of compounds (Clark, Gelman, and Lane 1985).

On the other hand, there is evidence that the development is much slower. Nicoladis (2003) presented results suggesting that children's subcategorization knowledge is not completed at the age of 3. In her experiment, children were presented with novel compounds (e.g., *dragon box*) and a set of pictures and asked to pick the picture that corresponds to the compound. Three-year-olds selected a picture showing a dragon next to a box rather than a box decorated with dragons more often than 4-year-olds. This suggests that 3-year-olds are still developing with regards to compound interpretations. Other studies present evidence that this process is not completed until well into the school years. Berko (1958) found that children between 4 and 7 years still had difficulties explaining the meaning of common compounds such as *birthday*. They often responded with a salient feature or function of the compound instead of an explanation that related modifier and head. In case of *birthday* they said it is called this way because one gets presents or eats cake. Only 2 percent of the children mentioned that it is a day (Berko 1958). While one might argue that a word like *birth* might not be fully understood by all children at this age, the results are in concordance with those by Parault, Schwanenflugel, and Haverback (2005), who compared interpretations of novel noun-noun compounds by 6- and 9-year-olds as well as adults. They found that children's interpretations, although quite adult-like, are nevertheless significantly different from those of adults. Striking were explanations that did not integrate the meanings of the constituents, but left them in an unconnected side-by-side status as in "a big magazine and a little book" as an explanation for *book magazine*.

These findings are in accordance with the assumption that children learn their first compounds as unstructured units and slowly develop knowledge of the roles of heads and modifiers as well as modifier-head relations. The seeming contradictions in the literature might indicate that children's

understanding of compounding differs from compound to compound. While they might be able to identify head and modifier of one compound, they might not be able to do the same for another compound. It might also be that children understand from early on that heads and modifiers are somehow related, but they appear to take a long time until their relation inferences become adult-like. These assumptions are in line with a usage-based theory of language acquisition, which assumes that children acquire linguistic constructions such as subject-verb-object or agent-action-patient on an item-by-item basis such as “I love you” and gradually generalize to more abstract patterns such as the subject-verb-object construct (e.g., Akhtar 1999; Goldberg 2006; Tomasello 2000, 2003).

In Krott and Nicoladis (2005) we investigated whether children’s understanding of the complex structure of a particular familiar compound is enhanced by the knowledge of constituent families. Are children more likely to parse a compound into head and modifier when they know other compounds with the same head or modifier than when they do not know other compounds? We asked English-speaking children between the ages 3 and 5 to explain to an alien puppet why we say compounds such as *chocolate cake*. We selected compounds that contained heads and modifiers with either large or small constituent families and confirmed the sizes of the constituent families by questionnaires given to the parents of the children. The results showed that the children were more likely to mention a constituent in their responses when they knew several other compounds with this constituent, i.e., when the constituent had a large constituent family (see Figure 6.2), both for modifier and head families. We confirmed this finding in an equivalent study with French noun-noun and noun-preposition-noun combinations (Nicoladis and Krott 2007). Together, these studies support the possibility that children’s understanding of specific familiar compounds relies on knowledge of similar compounds. In other words, the more exemplars there are that can form an analogical knowledge basis for the understanding of a compound, the better children understand the compound and the better they are in explaining its meaning.

In a recent study, we addressed the question whether the knowledge of constituent families also guides children’s interpretations of novel noun-noun compounds (Krott, Gagné, and Nicoladis, 2009). We asked adults and 4–5-year-olds to explain the meaning of novel compounds such as *dog shoes*. In accordance with previous research by Gagné and colleagues (Gagné 2001; Gagné and Shoben 1997), adults used their knowledge of relations in modifier families to infer modifier-head relations. For *dog shoes* they used their knowledge of other compounds with the modifier *dog* such as *dog house*,

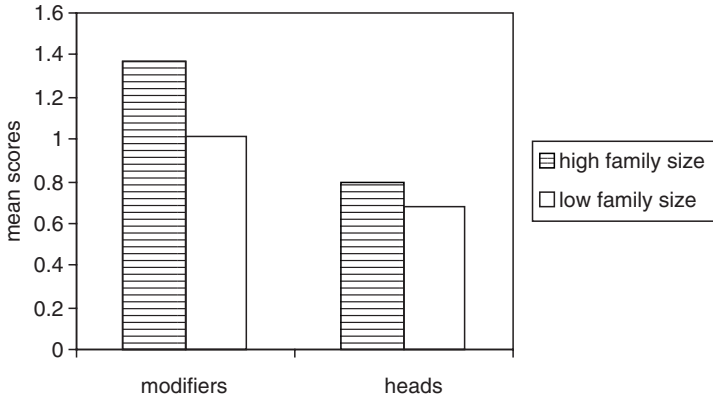


FIGURE 6.2 Children's average scores for modifiers and heads (max. 2) by family size (high versus low) in Krott and Nicoladis (2005).

dog biscuit or *dog leash*. Their interpretations were in line with the relational bias in modifier families, the relation FOR in case of *dog shoes*. Children's responses showed that they also used their knowledge of constituent families. However, they drew on their knowledge of relations in head families, i.e., other shoes such as *ballet shoes*, *snow shoes*, *horse shoe*. There was only weak evidence that they also used their knowledge of modifier families. That means that adults were influenced by their knowledge of how a particular modifier is used to create subcategories within different categories, while children were influenced by their knowledge of a particular category and the modifier relations in this category. It is not clear why children and adults should use different knowledge. One reason might be that children do not know much about possible modifications yet due to their limited compound vocabulary. As there is evidence that children at this age might still be developing their understanding of heads and modifiers, children might also focus on heads because identifying the category of the novel compound is the first step in understanding it. The latter might be linked to high task demands and the slow development of executive functions.

Ryder's (1994) research on adults' compound interpretations suggests that adults have access to a repertoire of analogical bases when interpreting compounds. They can choose between their knowledge of specific compounds, compound families, or more abstract schemas. The findings for compound acquisition, namely that children appear to know something about a very abstract level of compounds when they are as young as 2 years, while this knowledge does not appear to be fully developed yet when they are 4 or 5 or even later, might mean that young children develop analogical bases of

different levels of abstractness, but that the choice between these levels is not adult-like yet. Alternatively, as mentioned above, children's knowledge of a very abstract level of compounding might be an illusion. What looks like general knowledge about heads and modifiers might be knowledge about particular constituent families. Both explanations are in line with a usage-based theory of language acquisition (e.g., Goldberg 2006; Tomasello 2003).

6.5 Conclusion

We have seen that constituent families play an important role for noun-noun compounds and that this role is not limited to a specific aspect of processing, but appears to affect all types of aspects of compound processing. The analogical nature of this role is especially apparent for phenomena that are highly variable, i.e., that lack the systematicity of a rule-driven pattern. We have seen various types of evidence that constituent families provide an analogical basis for the production of compounds, in particular for the realization of phenomena appearing at constituent boundaries such as Japanese *rendaku* or German and Dutch interfixes as well as for stress assignment in English compounds. For the interpretation of novel noun-noun combinations, constituent families guide people's selection between various possible modifier-head relations. But even when compound processing is not characterized by choice and variability, constituent families play a role. We have seen their effect on the processing of familiar compounds, namely when familiar compounds are recognized in a lexical decision experiment or when they are judged as to whether or not they make sense. This suggests that the analogical basis of constituent families is not turned on or off depending on the task. When we process compounds, constituent families appear to be "active" regardless of the linguistic aim. They affect us when we create novel compounds as speakers and when we try to make sense of what somebody else says or writes.

It is remarkable that analogical compound processing is already in place in preschool children. Similar to adults, children use constituent families to discover internal structure in familiar compounds and to interpret noun-noun combinations that they have not encountered before. However, children seem to differ from adults in that they tend to focus on head families rather than modifier families. As pointed out, children's knowledge of possible modifications might not be developed yet. Future research will need to investigate whether this is indeed the case.

Throughout this chapter I have been assuming that constituent families are represented in the mental lexicon. We have seen that constituent families play an important role in very different domains of language processing. The question therefore arises whether the observed effects are based on a single lexical system that is involved in all these domains or whether they are based on two or more systems that are domain-specific, but structurally very similar. Morphological family effects that have been found for visual word recognition have been explained by overlapping semantic representations of family members (Bertram, Schreuder, and Baayen 2000; de Jong, Schreuder, and Baayen 2000; Schreuder and Baayen 1997). Thus, these effects are assumed to arise on a level of conceptual representations. The interpretation of novel compounds is likely to arise on the same level because it involves conceptual knowledge. Constituent family effects found for the production of novel compounds, however, are likely to arise on a level of morphophonological representation because these effects are all related to the form of the compounds. In the model of interfix selection in Krott, Schreuder, and Baayen (2002*b*) constituent family effects arise due to connections between morphophonological representations of compounds and interfixes. The voicing of obstruents in Japanese compounds and stress assignment in English compounds also involve morphophonological representations. The most likely scenario therefore is that constituent family effects, although very similar at first sight, originate from two structurally similar but nevertheless different subsystems of the lexicon, one situated at the level of conceptual representations, the other at the level of morphophonological representations. It is unclear whether constituent family effects at the morphophonological level might differ for written and oral processes because it is not clear yet whether morphophonological representations in the mental lexicon are domain-independent or not, i.e., whether there are different representations for written and oral comprehension and production (e.g., Caramazza 1997; Miozzo and Caramazza 2005). The results for interfixes rather suggest that morphophonological representations are domain-independent because interfixes occur in both oral and written production. Voicing in Japanese compounds and stress in English compounds, on the other hand, both concern only oral word production and therefore suggest domain-dependent analogical effects.

What also remains unclear is why for some phenomena it is the modifier family that plays the important role, while for other phenomena it is the head family. Taking together all findings, modifier effects seem to occur in languages such as German, Dutch, and Indonesian. These languages all have main stress on compound modifiers. However, it is unlikely that stress is the

driving factor. First, most of these studies presented compounds visually, i.e., without stress information. Second, for English, children revealed a stronger focus on head families, while adults revealed a stronger focus on modifier families for the exact same compounds. The focus on one or the other constituent is more likely due to distributional patterns in the language. For instance, as had been shown in simulation studies with TiMBL and AML, interfixes in Dutch or German compounds are better predicted by modifiers than heads and Japanese *rendaku* is better predicted by heads than modifiers. Speakers appear to be sensitive to this information and to make use of it.

In sum, we have seen how an entire class of words across languages and language families can be governed by analogy. It is likely that analogy is not restricted to noun-noun compounds, but that it plays an important role for other areas of morphology as well. It is therefore not at all unlikely that analogy underlies regularities that appear to be governed by rules.

Acknowledgment

The author would like to thank Antje Meyer and two anonymous reviewers for their comments on an earlier version of this chapter and James Myers for bringing the studies on Chinese constituent families to her attention.

Morphological analogy: Only a beginning

John Goldsmith

All reasoning is search and casting about, and requires pains and application.

John Locke, *An Essay Concerning Human Understanding* (1975 [1690])

7.1 Introduction

The perspective that I will describe in this paper is the result of some work over the last ten years or so aimed at building an automatic morphological analyzer—that is, an explicit algorithm that takes natural language text as its input, and produces the morphological structure of the text as its output.¹ The main conclusion, as far as analogy is concerned, is that formal notions that correspond very naturally to the traditional notion of analogy are useful and important as part of a boot-strapping heuristic for the discovery of morphological structure, but it is necessary to develop a refined quantitative model in order to find the kind of articulated linguistic structures that are to be found in natural languages.

I take the perspective that the three principal tasks (we could call them the *first* three tasks) of someone who wishes to develop a theory of morphology that applies to natural languages is to develop an account for (1) the segmentation of words into morphs; (2) the description of a grammar to generate words, on the basis of the morphs, among other things; and (3) the labeling of morphs, in two different ways: (a) a labeling that indicates which morphs are different realizations of the same morpheme, and (b) a labeling that indicates the morphosyntactic feature representation of each morpheme. Of these three, I will focus on the first two, and of the first two, I will emphasize the first. I underscore this because if we were historians of linguistics in the future

¹ I am grateful to Juliette and Jim Blevins, to Susan Rizzo, and to anonymous referees for comments on the original version of this chapter.

looking back at what questions were the focus of discussion in the first decade of the twenty-first century, it would appear that the first question must have been settled, in view of how little discussion there is of it.² I mean very simply, how do we justify the statement (for example) that *books* is composed of two morphs, *book* and *s*, while *tax* is not? One of the reasons that the problem of segmentation is interesting is that we cannot call upon the resources within generative grammar that most of us are familiar with, and have grown dependent upon—which is to say, appeal to substance in an innate Universal Grammar. There is no plausible account of how speakers of English learn that “*ing*” is a suffix, while speakers of Swahili learn that “*an*” is a suffix, that appeals to a small list of discrete parameters, each with a small number of settings.³ In fact, from a certain point of view, this is one of the reasons why the study of morphology so interesting: there is so much that must be learned.

I will begin with a discussion of the computational problem of *word* segmentation—that is, the problem of dividing a long string of symbols into words, with no prior knowledge of the words of the language. This is one of the problems that any child language learner faces. We will see that a large part of the difficulty that we run into when we tackle this problem derives from the importance of having a good model of morphology, without which all of our efforts to learn words would be in severe trouble. Rather than trying to solve both problems at the same time (the problem of word segmentation, and the problem of morphology induction), we will turn

² There is a perspective on word structure, articulated notably by Rajendra Singh and Sylvain Neuvel (Neuvel and Singh (2001), Neuvel and Fulop (2002)), which denies the existence of morphs and the internal segmentation of words. While I appreciate the force of their arguments, it seems to me that the same arguments against the decomposition of words into morphs holds, with essentially the same degree of conviction, against dividing sentences up into words—there are unclear cases, there is semantic noncompositionality in quite a few cases, and so on. But at the same time, it seems to me that linguists have to agree that concatenation is the preferred formal operation in both morphology and syntax, and the focus on segmentation into words and morphs can be understood as no more and no less than a consequence of that preference.

³ There is a tradition of no great antiquity in linguistic theory of seeing the adult grammar as a collection of objects selected from a fixed, universal inventory of objects, rather than as an algebraic representation of some sort whose length is in principle unbounded. The first explicit mention of this, as far as I know, is found in David Stampe’s work on natural phonology in the early 1970s (see Stampe (1980) [1972]), followed by Daniel Dinnesen’s atomic phonology (see Dinnsen (1979)); the strategy was adopted in Chomsky’s principles and parameters at the end of the 1970s, and it has never left the charts since then. It gained renewed vigor with the rise of optimality theory in the 1990s. Its appeal is no doubt due to the pious hope it has been known to inspire that the problem of language learning may turn out to be trivial, because the differences between languages will amount to a small number of bits of information. I find this sad, in part because, if we can’t count on linguists to tell the world about the richness and variety found across humanity’s languages, there is no one else to do it. It’s doubly sad, in that even if it *were* the case that learning a language could be modeled as being much like selecting a set of, say, 50 items out of a universal set of 1,000, we would still need to do some heavy lifting to produce an account of learning; since there are some $\frac{1000!}{50! \cdot 950!}$ ways to do that, the fact that this is a finite number is not much consolation. I will return to this in the conclusion.

specifically to the task of discovering the morphology of a language with no prior knowledge of the morphology, but with prior knowledge of where word boundaries are (as if we had already solved the word segmentation problem), and discuss the role that analogy plays in this latter task. Naturally, we would like to merge these two tasks, and present an algorithm that takes an unsegmented segment stream as input and produces both a word list and a morphology; we are not yet able to accomplish that (though I suspect we have the tools at our disposal now to tackle that problem). I would like to emphasize, however, that the materials on which we base our experiments are not prepared corpora or toy data; they are in every case natural materials from natural languages.

There is a more general point behind my account as well, which deserves at the very least a brief presentation before we settle into a discussion of a specific problem. It is this: the present paper assumes that we can specify a scientific goal for linguistics which is independent of psychology, and which depends only on computational considerations. Being independent of psychology, it does *not* presume to tell psychologists what conclusions they will or should reach in their exploration of the human mind and brain, nor does it depend on those explorations. Its premise is very simple: given a particular corpus from a language (that is, a finite sample, which can be as little as a few thousand words, or as large as the internet as of some moment in time, like today), the goal is to find the best grammar (or set of grammars) that accounts for that data. This suggestion is only as useful as our ability to explicate what it means for a grammar G to account for a set of data, or corpus, C , and we will define this as the probability of the grammar G , given the data C ; and we will see below that by this we shall have meant the grammar G such that its probability (based on its form) multiplied by the probability that G assigns to C , is the greatest. How such a view is possible and reasonable will become clearer shortly.⁴

Before proceeding any further, I would like to say what I mean by analogy in morphology. Unless specified otherwise, I will assume that our goal is to analyze the internal structure of words, and also that we actually know where words begin and end in the sound (or letter) stream of the language we happen to be looking at. In fact, I will assume that our problem is to find internal structure when presented with a word list in a language. In traditional terms, *book* : *books* :: *dog* : *dogs* would constitute an analogy; so would *jump* : *jumped* : *jumping* :: *walk* : *walked* : *walking*. A more perspicuous way to look at this sort of analogy is as in (1), which we call a “signature”; a computer

⁴ This notion is also presented, in greater detail, in Goldsmith (2007).

scientist would prefer to represent the same data as in (2), which he would call a representation of a finite state automaton (FSA).

$$\left\{ \begin{array}{l} \text{walk} \\ \text{jump} \end{array} \right\} \left\{ \begin{array}{l} \emptyset \\ \text{ed} \\ \text{ing} \end{array} \right\} \quad (1)$$



But before we talk about morphological analysis, let us turn first to the problem of word segmentation.

7.2 The problem of word segmentation

In the mid-1990s, Michael Brent and Carl de Marcken (both graduate students in computer science at the time working with Robert Berwick at MIT) developed computational methods for inferring word boundaries in a continuous stream of discrete symbols, relying on Minimum Description Length (or *MDL*) analysis (Brent (1999), de Marcken (1996), Rissanen (1989)). Their projects could be interpreted (as they did interpret them) as representing an idealization of how a child can learn the words of a language when exposed only to a stream of phonemes. This is the *word segmentation problem*: how to find words in a larger stream of symbols. Now, there are two fundamentally different approaches that one could take in dealing with the word segmentation problem (and one could certainly adopt both approaches, since they are not incompatible): one can either focus on finding the *boundaries* between words, or focus on finding words *themselves* in the stream, the sequences of recurring symbol strings, and inferring the boundaries from knowledge of the words. I think that there is a widespread (and natural) tendency to feel that the first of these two methods (finding cues in the signal that show where the boundaries between words are) is the more appealing way to approach the problem, perhaps on the grounds that you cannot take the second approach without engaging in some kind of inappropriate circular reasoning. This intuition is probably encouraged, as well, by the observation that in a good number of European languages, there are relatively straightforward superficial phonological cues to mark the delimitations between words, such as can be found in words in which the initial syllable is regularly stressed (as in Finnish, and as was once the case in German), or in which the penultimate syllable is stressed.⁵

⁵ As an aside, I would mention my belief that this approach is hopeless as a general solution to the problem of word segmentation. The reason for this pessimistic view is that the difference in

The second approach, as I noted, is to say that we will *first* find the words in the signal, and *then* divide the signal up into words in the most likely way based on that knowledge of the words, along with the assumption that the speech signal can be partitioned without overlap into a succession of words. But how can this kind of learning be done?

I will give a brief summary here of the Brent-de Marcken approach to answering this question, based on MDL modeling. My account leans more heavily on de Marcken's specific approach than on Brent's, but it is a simplification of both, and the reader who would like to learn more is strongly advised to read the original works.

Minimum description length modeling was first developed by the Finnish-American statistician, Jorma Rissanen, notably in a book published in 1989 (Rissanen (1989)). The question he is concerned with is not specifically linguistic at all. It is simply this: given a body of data, how can we be sure to extract all and only the regularities that inhere in the data? We want to fit the model to the data, or the data to a model, and we want neither to overfit nor to underfit. Underfitting would mean failing to extract some significant regularity in the data; overfitting would mean misinterpreting something that was, in some sense, accidentally true of the data which was sampled, but would not be true of a larger sample from the same source.

Rissanen's approach is inherently probabilistic in two ways. To explain what these ways are, I shall discuss the problem of word segmentation in particular, even though Rissanen's approach is very general and was not developed with linguistic problems in mind. The first way in which the MDL approach is probabilistic is that an MDL analysis is a model (or grammar) that assigns a probability to every conceivable string of phonemes (or letters, if we are working with a language sample from a written source). This is a stringent condition: a probabilistic model is by definition one which assigns a non-negative number to every possible input, in such a fashion that the grand total of the probabilities adds up to 1.0—and this must be true even if the set of possible inputs is infinite (which is virtually always the case). Probability is thus *not* a measure of something like uncertainty or randomness; if anything, imposing the condition that the model be probabilistic imposes a very tight

probabilities that such approaches can assign to cuts in different places in a sound stream are far too small to allow a successful overall division of the stream to be accomplished in a local way, that is, based only on local information. The problem can only be solved by maximizing the probability of a parse over the longer string, which allows us to take into account the probabilities of the hypothesized words, as well as the conditional probabilities of the hypothesized words. To put this in a slightly different way, in order to segment a stream into words, it is not sufficient to have a model that predicts the phonetics of the word boundaries; one must also have a language model, assigning a probability to the sequence of hypothesized words. The interested reader can find a survey of much of the material on segmentation in Goldsmith (2009). See also Roark and Sproat 2007.

overall constraint on the system as a whole. In the language of probability, we are required to specify ahead of time a sample space and a distribution over that sample space; the distribution is essentially a function that maps a member of the sample space (or a subset of the members of the sample space) to a real number, in such a way that the whole sample space maps to 1.0.

The second way in which an MDL analysis is probabilistic is more abstract. We set a condition that the grammars *themselves* are the subject of a probability distribution; which is to say, every possible grammar is assigned a probability (a non-negative real number), subject to the condition that these probabilities sum to 1.0—and this must be true even if the set of possible grammars is infinite (which is virtually always the case). The reader may note that this condition puts MDL within the broader context of approaches which includes Bayesian approaches to modeling; MDL puts the priority on the quantitative notion of encoding, both regarding the data and the grammar, but there is an overall commonality from a distant enough perspective.

Although it may not sound like it at first, this second condition is very similar in spirit to Chomsky's view of grammar selection in early generative grammar (that is, in classical generative grammar (Chomsky (1975 [1955])), which in the late 1970s many generative grammarians abandoned—after little discussion—in favor of the principles and parameters approach (Chomsky and Lasnik (1977))). According to this perspective, the primary goal of linguistic theory is to make explicit a formalism for grammar writing, but not just *any* formalism. The goal was a formalism with which predictions (or, more modestly, claims) could be made as to which grammar was correct among a set of grammars all consistent with the given data; those predictions would be based purely on the length of the grammar in the some-day-to-be-discovered formalism.

MDL employs a few simple ideas to assign a probability to a (potentially infinite) set of grammars, and we should at least sketch these ideas. Perhaps the most important is what is known as Kraft's inequality. Kraft's inequality holds for uniquely decodable codes, but we will consider (as does most MDL modeling) a special case of that—those codes which are said to respect the prefix condition. The term *coding* here should simply be interpreted as meaning something like *formalized as a grammar*, and in general we want to consider the class of all grammars that are permitted by a certain formalism. The prefix condition sounds innocuous: it says that there are no two grammars (G and H , say) which have the property that H equals all of G plus some additional material (as computer scientists put it: there are no two grammars G and H such that G is a prefix of H —but remember that computer scientists just use “prefix” to mean a substring that starts at the beginning of some other string). Another way to put it is this: when you are reading a grammar, you

know when you reach the end of it. (The condition seems innocuous, but its consequences are major, for reasons that we will not go into here.)

Kraft's inequality says that if a set of strings (here, grammars) does indeed respect the prefix condition, then we can assign a probability to each string (grammar) S equal to $2^{-\text{length}(S)}$. Why the number 2 here? I have assumed (as computer scientists tend to) that we encode the grammar using strictly binary encodings, the way a computer does, using only 0's and 1's. If we want to use a vocabulary like the Latin alphabet, then the base is going to be 26—or more likely 27, if we include a punctuation symbol, like space,⁶ and so below I will replace “2” by “27.” If the length of a grammar is 100 0's and 1's, then we assign it a probability of $\frac{1}{2^{100}}$; if it's 100 *letters*, then we assign it a probability of $\frac{1}{27^{100}}$. Unless we're very careful with our assignment of lengths, this quantity (based solely on grammar length) will sum to a finite number less than 1 (call it k); and then, to turn these numbers into true probabilities, we divide each of them by k , so that the sum totals 1.0.

In short, with a very mild condition (the prefix condition) imposed, we can easily specify a natural probability distribution over the infinite class of grammars, according to which a shorter grammar is a more probable grammar. In fact, if grammar G has length g , and grammar H has length h , then the ratio of their probabilities is simply $2^{(g-h)}$ if binary encoding is used, and $27^{(g-h)}$ if the Latin alphabet is employed.

Now we take two further steps. The first involves Bayes' rule, which is nothing more than an algebraic restatement of the definition of conditional probability. The second involves the assumption that there is a single correct answer to our question.

Bayes' rule says that (in the case that we are considering) the probability of a grammar, given our corpus, is closely related to the probability of the grammar, given the corpus, as follows:

$$\text{pr}(G|D) = \frac{\text{pr}(D|G)\text{pr}(G)}{\text{pr}(D)} \quad (3)$$

The left-hand side refers to the probability of a grammar G , given the data D at hand (i.e., the corpus), while the right-hand side is the product of the probability of the corpus assigned by the grammar G , times the probability of the grammar, divided by the probability of the data. Since our goal is to find

⁶ There is a fine line here between clarity of exposition and accuracy of modeling. In general, we *don't* want to use special boundary symbols to demarcate the ends of representations, because this is typically a wasteful and inefficient way of marking boundaries; an encoding which respects the prefix condition is better. But ease of exposition will sometimes trump formal niceties in this chapter.

the grammar whose probability is the greatest (given the data at hand, and what else do we have other than the data at hand?), we can interpret (3) to mean: find the grammar G for which this quantity is the greatest. The denominator, $pr(D)$, is perhaps the hardest to compute, but we do not in fact need to calculate it, because it is a constant. Since we have just finished discussing how to calculate the probability of the grammar G , based on its length, calculating $pr(G)$ is not a problem. And calculating $pr(D|G)$ is not a problem, either, since we have assumed from the start that our model is probabilistic, which is to say, that it assigns a probability to every conceivable corpus. So in the end, our task simply boils down to this: find the probabilistic grammar G such that the probability of the corpus, given the grammar, times the probability of the grammar itself, is the greatest.

Brent's and de Marcken's insight was that the method that we have just described could be applied to the problem of word segmentation and lexicon induction. We need to do three things: first, figure out how a lexicon (with its probability) actually assigns a probability to any corpus; second, figure out how to associate a lexicon with a length, so that we can in turn assign it a probability; and third, figure out how to actually come up with a candidate lexicon, along with probabilities assigned to each word in the lexicon. It turns out that none of these is too difficult, at least as a first approximation.

First, how do we assign a probability to a corpus D , given a probabilistic lexicon? We need to take into consideration the fact that there will, generally speaking, be many ways of parsing a corpus up into words. If all we know about English is its words (and nothing about syntax, meaning, and so on), then a string like: THISMEANSTHAT that can be divided up in many ways. There is THIS-MEANS-THAT, but then (since every individual letter can be used as an individual word in languages, in general), there is also THIS-ME-AN-S-THAT, and T-HIS-ME-AN-S-T-HAT, and many others. So first of all, we make the assumption that only one parse of a given corpus is actually correct,⁷ and that the parse that is assigned the highest probability by our corpus is the correct one. And the probability assigned to a given parse is defined as the product of two factors: the first is the probability that the corpus has exactly as many words in it as the parse has pieces, while the second is the product of the probabilities of all of the words in the parse. In

⁷ That this assumption is a bit too strong is illustrated by the ambiguity of phrases like "cookmea-napplesauce", which has perhaps two reasonable parses: *cook me an apple sauce* and *cook mean apple sauce*. The reader is invited to construct similar examples in other languages.

the case of THIS-MEANS-THAT, the probability of that parse is equal to the probability that a string has three words in it, times the product of the probabilities of each of the three words *this*, *means*, and *that*.⁸

Second, what is a lexicon's length? If we define a lexicon as a concatenation of words, then as long as we separate each of the words by a space, the words satisfy the conditions for Kraft's inequality, and we can assign a (prior) probability to a lexicon equal to 1 divided by 27 raised to the power of the length of the lexicon, in letters: $\frac{1}{27^{\text{length}(\text{lexicon})}}$.

Third, how do we *find* a lexicon, given a corpus? We proceed in a bottom-up fashion, assuming initially that the lexicon consists of all the letters of the corpus. Then we iteratively repeat the following process: we look at all "words" that appear next to each other in the corpus, and pick the most frequent such pair. (Initially, this may be T-H in the case of a written corpus of English, since our initial assumption is that the words of the lexicon are the letters of the language). We use our MDL criterion to decide whether to declare that T-H is really a word TH. Our MDL criterion is simply this: does the expression described in (3) increase when we add our candidate to the grammar? Does the probability of the corpus increase enough by the addition of TH (for example) to offset the decrease in probability of the lexicon that comes about from increasing its length (from 26 real members to 27, the alphabet plus TH)? If so, then we include the new member; if not, we leave the grammar as it is and try some different candidates. This process stops when there are no neighboring chunks in the corpus whose addition would increase the overall probability of the corpus.⁹

There is one more step that we need to take to appreciate the beauty of Rissanen's MDL framework. If we take the logarithm of both sides of equation (3) and multiply these two expressions by -1 , we obtain the following quantity: $-\log pr(D|G) - \text{length}(G) + \log pr(D)$. The third term is a constant. However, the first term has a very real significance: it is called the *optimal compressed length of the data*, and the second term also has a real significance: it is, quite simply, the length of the grammar, which we use in order to evaluate how well the grammar succeeds at being a compact formulation. The first term, the optimal compressed length of the data, given the model, is a well-understood quantity expressing how well the model does at extracting generalizations from the data. Thus the task of finding the grammar that *minimizes* this quantity (*minimizes* instead of *maximizes* because we

⁸ There are several ways to establish a reasonable distribution over number of words in sentence, but they do not bear on our discussion here.

⁹ See the Appendix.

multiplied it by -1 , and the logarithm function is monotonic increasing) is equivalent to finding the most probable grammar, given the data at hand.

We intend by this to mean what was suggested above: there are no constraints on the forms of possible grammars, above and beyond the condition that they be programs for a Turing machine, and thus are algorithms.¹⁰ This means that the purpose of linguistic theory is to serve as a set of heuristics to help the linguistic scientist come up with a tight, snug grammar, given a set of data. MDL can determine which of a set of grammars is the best one, given the data; no feasible process can search all possible grammars, so there is no guarantee that another linguist will not come along tomorrow with a *better* grammar for the data. But it will be *truly* better, better as far as the length of its Turing machine program is concerned. We know that there is a best analysis (up to the unlikely possibility that two or more grammars have (along with the data) an equal description length), because the minimum description length will be some positive number *less* than the description length provided by the (dumb) grammar consisting of exactly the corpus with no internal structure (along with some reasonable closure conditions).

7.3 Success with word discovery?

How well does this method work? Anyone who has worked with corpora knows that, to some extent, an answer to this question depends heavily on the corpus used for training and for testing. In the case at hand, there is no training corpus as such; the input to the algorithm is a long string that has no indication of word boundaries, and the output is a guess (or prediction) as to where the word boundaries are, or should be. In view of the fact that the system has no prior knowledge of the language, the results are in some respects very impressive, but at the same time, when we look at the results with the eyes of a linguist, we quickly see some linguistically things that have gone awry.

In Figure 7.1 is the beginning of a passage from the first 100,000 words of the Brown corpus and Figure 7.2 is the beginning of a similar passage from a Portuguese document.

Three things jump out when we look at these results. First, there are many errors caused by the algorithm finding “pieces” that are too small, such as *produc-ed*: it seems as if the system is finding morphemes in this case, while in

¹⁰ The point may be purely terminological, but I would argue that the position I am describing clearly falls under the definition of generative grammar, at least as it was considered in Chomsky (1975) [1955]; algorithmic complexity is the simplicity metric utilized.

The Fulton County Grand Jury's aid Friday an investigation of Atlanta's recent primary election produced no evidence that irregularities took place. The jury further said in term - end present ment s that the City Executive Committee, which had over - all charge of the election, does serve the public and than k so the City of Atlanta for the manner in which the election was conducted.

FIGURE 7.1 The first sentences of the Brown Corpus

De muitos outros recursos da floresta, não apenas folhas, flores era íz esmas também de se mente se da casa de árvores retiram produtos medicinais como quais se habituaram nas uas o lição e nos seus sonhos a enfrenta ra s do enças que hoje coma chega da dos branco s começa ma trata r como re médi os da indústria urbana - e que muitas vezes não produz em e feito.

FIGURE 7.2 The first sentences of a Portuguese document

other cases it is finding words. Second, in some cases the algorithm finds pieces that are too big: they are “pieces” like *forthe* which occur together often enough in English that the algorithm erroneously decides that the language treats them as a word. Third, there are far too many single letter words: we need a prior probability for word length that makes the probability of one-letter words much lower.

We will focus here on just the first of these points. Why should the system find morphemes rather than words some of the time? The answer is perhaps obvious: the system that we are considering is nothing more than a lexicon, bereft of any ability to find structure in the data other than frequency of appearance of strings of various lengths. There is no ability built into the system to see relationships between words, nor any ability to see that words may enter into relationships with the words around them. We need to add linguistic structure to this approach, then. And that is what we turn to now.

7.4 The *Linguistica* project

I have been working since 1997, along with Colin Sprague, Yu Hu, and Aris Xanthos, on the development of a software package, *Linguistica*, whose primary goal is the automatic inference of morphological structure on the basis of an unmodified sample corpus from a real language, and whose

method is MDL as we have described it in this chapter; see <http://linguistica.uchicago.edu>¹¹

A big, and I would say controversial, assumption made by the *Linguistica* project is that meaning can be ignored in the process of inferring or inducing the morphological structure of a word or a language. The fact is, the procedures we have explored make little or no reference to meaning. Any successes that we achieve can be interpreted as showing that reference to meaning is not *necessary*, but we certainly cannot infer that human language learners do not use meaning in their search to discover language structure. It is natural to interpret our project as an effort to figure out, from a linguistic point of view, exactly *where* a learner, one who has access neither to a rich innate component nor to the meaning of utterances, will fail.

In some ways, the work that I am describing could be viewed as a neo-Harrisian program, in the sense that Zellig Harris believed, and argued, that the goal of linguistic theory was to develop an autonomous linguistic method of analyzing linguistic data, in which the overall complexity of the grammar was the character that the linguist would use in order to evaluate competing analyses, and in which the linguist was, in the final analysis, more interested in the methods of analysis than in the analysis of any particular language.¹² As long as we are clear what we mean by the term *discovery procedure*, it would be fair to say that this work aims at developing a discovery procedure for morphology. While it does not propose a simple step-by-step process for this end, it does propose something so close to an algorithm as to be indistinguishable from a computer program—which is why it has been relatively easy to encode the proposals as computer code which can be tested against small and large natural language corpora.

7.5 MDL, grammar simplicity, and analogy

One way to summarize what MDL methods have in common is to say that they seek to extract redundancy in the data. In the case of word segmentation, the redundancy is the reappearance of the same substrings on many occasions, while in the case of morpheme discovery, it is the reappearance of morphemes under quite particular and restricted conditions. What I will describe here is a considerable simplification of the model as it actually works, and the reader can find detailed discussion in Goldsmith (2001, 2006). As we saw above, the prior probability that is assigned to a grammar is based entirely on its

¹¹ See Goldsmith (2000, 2001, 2006).

¹² See Goldsmith (2005) for a recent discussion.

length, quite literally, and hence any redundancy in the formulation of a grammar leads to a heavy cost paid by the grammar, in terms of the lowering of the probability assigned to it. Conversely, a grammar which has been shortened by the elimination of redundancy is assigned a considerably higher probability. And, as we will see, analogy is one essential way in which redundancy can be discovered by the language learner.

The basic idea is this: when sets of words can be broken up into two pieces in precisely parallel ways (as in the signature shown in (1), repeated here as (4)), we can extract measurable redundancies. Here, we have taken the six words *jump*, *jumped*, *jumping*, *walk*, *walked*, and *walking*, and observed that there is a pattern consisting of two distinct stems, and three distinct suffixes, and all combinations of stem and suffix appear in our data set.

$$\left\{ \begin{array}{l} \text{walk} \\ \text{jump} \end{array} \right\} \left\{ \begin{array}{l} \emptyset \\ \text{ed} \\ \text{ing} \end{array} \right\} \quad (4)$$

Before any such analysis, we were responsible for encoding all the letters of the six words, which comes to forty letters (including a final space or word boundary), while after we extract the regularity, only sixteen letters need to be specified (again, counting a boundary symbol along with each suffix).

In somewhat more useful—that is, generalizable—terminology, we can describe this data with a finite state automaton (FSA), as in (2), repeated here as (5).



To encode this, we need a formal method for describing the three states and their transitions, and then we need to label each transition edge; we have already seen a simple (and, as it turns out, overly simple) way of measuring the complexity of the labels, which was by counting the number of symbols. We will ignore the computation of the complexity of the FSA itself; it is very simple from a technical point of view.¹³

¹³ Each FSA consists of a set of pointers to nodes, along with labels that are themselves pointers to strings. A maximum likelihood model provides probabilities in each of those two domains; the complexity of the overall FSA is the sum of the inverse log probabilities of all of the pointers in the representation.

This overall system can then naturally be regarded as a device capable of expressing morphological analogies of the *book : books :: dog : dogs* sort. How does it operate in practice? Does it work to find real linguistic morphological regularities?

The answer, in a nutshell, is this: we can find patterns, locally and in the small; but a very large proportion of them are spurious (that is to say, linguistically wrong and irrelevant) unless they participate in larger patterns of the language as a whole. An example of a linguistically real discovery is as in (4) or (5), and a spurious example is as in (6), which captures the nongeneralization inherent in the words *change, changed, charge, charged*, or (7), which captures the nongeneralization inherent in the words *class, cotton, glass, gotten* (and I could offer dozens of examples of this sort from any language of which we have a few thousand words in computer-readable form: it was not I, of course, who discovered these patterns, but rather an over-eager analogy-seeking computer program):

$$cha \left\{ \begin{array}{c} n \\ r \end{array} \right\} ge \left\{ \begin{array}{c} \emptyset \\ d \end{array} \right\} \quad (6)$$

$$\left\{ \begin{array}{c} c \\ g \end{array} \right\} \left\{ \begin{array}{c} lass \\ otten \end{array} \right\} \quad (7)$$

What is wrong with the spurious generalizations in (6) and (7) is that the proposed morphemes do not appear outside of this generalization, more generally in the language. Analogy, as we see it here, is an excellent and important source of hypotheses, but it is not more than that. We need to develop means (and, it appears, largely formal means) to evaluate the hypotheses suggested by analogies.

The use of Minimum Description Length analysis provides at least a part of the response to this need, and it sheds some interesting light on the role played by information theory in linguistic description. Embedded within the work cited above by de Marcken is the key insight formalized by the use of information-theoretic formalisms—namely, that reuse of a grammatical object (such as a morpheme, a context, or anything else) is the best kind of evidence we can have of the linguistic reality of the object. What makes the *n, r* pairing in (6) linguistically irrelevant is the small number of times it is found in the linguistic analysis of English—unlike the \emptyset, d pairing, but like the *c, g* pairing in (7).

But this should not lead us to thinking that we simply need to count occurrences and look for some magic threshold count, because information theory provides a much better method for understanding what is at play. The key point is this: the edges in the finite state automaton in (5) should be understood not as being labeled with strings of phonemes, but rather as being labeled by pointers to morphemes in a separate inventory of morpheme spell-outs. This simple formal decision has two consequences. The first is a consequence that comes from information theory: the complexity (in quantifiable bits) of a pointer to a morpheme is directly controlled by the frequency with which a morpheme is used throughout the grammar. The second is that we arrive at a natural understanding of the view, famously voiced by Meillet, that language is a system in which everything is interconnected.¹⁴

The decision to label edges of a morphology with pointers rather than phonic substance makes strong predictions: strong enough to build a program that figures out the structure by itself, without human oversight. *Linguistica* discovers affixes by seeking robust clusters of stems and affixes, such as the large set of stems in English that take exactly the suffixes \emptyset , *ed*, *ing*, *s*. But what of stems that occur with an idiosyncratic set of affixes, a set of affixes shared by no other stem? Consider the examples in (8) and (9).

$$act \left\{ \begin{array}{c} \emptyset \\ ed \\ s \\ ion \end{array} \right\} \quad (8)$$

$$car \left\{ \begin{array}{c} d \\ e \\ l \\ p \end{array} \right\} \quad (9)$$

Each of these signatures is an example of a stem that appears with exactly four suffixes in a pattern shared by no other stem in a particular corpus. But the information-theoretic cost of building a pattern with the suffixes in (8) is much less than that of building the pattern shown in (9)—not because of the number of letters (phonemes) in each case, but rather because */l/* and */p/* are both rare affixes in English (note: *affixes*, not phonemes). An affix that occurs

¹⁴ In particular, “Comme pour tout autre langage, les différentes parties du système linguistique indo-européen forment un ensemble où tout se tient et dont il importe avant tout de bien comprendre le rigoureux enchaînement” (Meillet (1915) p. x).

on one word in a lexicon of 20,000 words will “cost” approximately $\log_2 20,000$ bits (about 14 bits), while a suffix that occurs on 1,000 words will cost about 4 bits—a very large difference, in the event; and the cost of positing /l/ and /p/ as affixes outweighs the gain saved by positing /car/ as a stem in (9). The same is not true of the case in (8), where the cost of building a subgeneralization to deal with the words based on the stem /act/ is much cheaper, because all of the observed suffixes are cheap, in an information-theoretic sense: they are independently used enough throughout the grammar that using them additionally in the creation of a new generalization costs the grammar very little. This implicit “thought process” is easy to formalize and to embed within an automatic morphological analyzer.

In Table 7.1, I have given some data from a sequence of steps of learning the morphology of the first 100,016 words of the Brown Corpus.

The first row in Table 7.1 shows the length of the “trivial” morphology at the beginning: it expresses the phonological cost (so to speak) of listing all 13,005 distinct words without any analysis: all words are stems, no stems are analyzed (we speak of “cost” to underscore the fact that we try to minimize this quantity). Row 2 shows the result of a relatively conservative effort to find signatures with several stems and several affixes, and we see that the information stored in the analyzed stems is now 53,835, while the information that we have taken away from the unanalyzed stems is greater: it is the difference between 486,295 and 390,160 (or 96,135). The additional infrastructure (affixes plus signatures) to accomplish this cost 1,220 + 22,793 (=24,013), for a total cost of 53,835 + 24,013 = 77,848. This cost (77,848) is much less than what was saved (96,135); the difference is 96,135 – 77,848 = 18,287. (Against this gain must be reckoned a slight decrease in the probability computed for the corpus.)

In the third, fourth, and fifth rows, we see the result of extending the discovery of signatures, stems, and affixes accomplished on the first pass to

TABLE 7.1 Description Length of morphology evolution during learning

Steps	Total	Unanalyzed stems	Analyzed stems	Affixes	Signatures
1. Before analysis	486,295	486,295	0	0	0
2. Bootstrap heuristic	468,008	390,160	53,835	1,220	22,793
3. Extend known stems and affixes	456,256	377,635	58,835	1,220	23,566
4. Find new signatures	434,179	320,405	74,440	1750	37,584
5. Find singleton signatures	429,225	235,390	128,830	1710	63,295

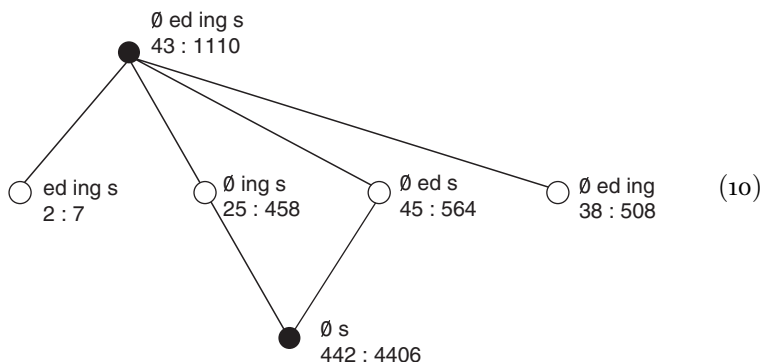
analyze words that were not initially analyzable. These are words for which the simple analogies of the first step were insufficient to uncover them, which include the discovery of patterns as in (8) and the rejection of those like in (9).

The algorithms explored in Goldsmith (2006) are remarkably good at discovering morphemes and morphological structure in a language with a complexity comparable to that of English. In the next sections, I will focus not so much on what they get right (which is better covered in the papers I have cited) but rather on where the challenges (some of them quite daunting) appear to be.¹⁵

7.6 The challenging of “collapsing” cases

Consider once more the case of English, where stems can be followed by a rather small set of affixes: verbs by $\{\emptyset, ed, ing, s\}$, nouns by $\{\emptyset, s\}$, adjectives by $\{\emptyset, er, est\}$. In even a modest-sized corpus, we will find a large number of stems that appear with all of their suffixes inside the corpus. But in addition, we will find a good number of stems that only appear with a subset of their possible suffixes. In the simplest case, this is due to the fact that the stem did not appear very often in the corpus. This is illustrated in (10), where each node represents one of the signatures, or small FSAs, that we have considered, and it is labeled with its set of suffixes. Below the label are two numbers: the first indicates the number of distinct stems that occurred in the corpus with this set of suffixes, and the second indicates the total number of words that occurred with these stems and suffixes. The two filled nodes are the “saturated” ones in which, from a linguistic point of view, all the suffixes that could have appeared have appeared. The node on the top row has four suffixes; those on the middle row have three suffixes, each a subset of those of the node on the top row; and the node on the bottom row has two suffixes, a subset of the two nodes from which it hangs on the middle row.

¹⁵ A reviewer of this chapter noted that “work on morphological processing (e.g. Baayen and Moscoso del Prado Martín (2005); Hay and Baayen (2005)) and [other work by Ernestus and Baayen]) suggests analogical relations are sensitive to semantic similarity, phonetic similarity, frequency effects, and more”. The information-theoretic models of the sort discussed in the present chapter give a firm theoretical foundation for why frequency effects are found; the reason is that information links in a grammar contribute a measurable amount to the complexity of the system, and that amount is equal to the reciprocal of the logarithm of the element being linked to. In the morphological analyses that we have studied in the *Linguistica* project, phonetic similarity has never emerged as a factor which, if integrated, would allow for superior performance. The relevance of semantic information is a difficult question; while I believe that it is relevant and could potentially improve performance in many cases, it is not easy to integrate meaning into a learning algorithm in a way that does not beg the question of learnability by building in too much information and treating that information as if it had been observable.



We need a method that determines that the white nodes in (10) are only partial generalizations, while the filled nodes are complete. To be sure, I have expressed this in categorical terms, when it is clear (or it becomes clear, when we look at more data) that the distinction is a soft one, rather than a hard one—but discussion of this point would lead us afield. I will return below to this question in the context of a language like Swahili, where it becomes even more pressing. To rephrase the problem, we can ask, when we have two signatures that are partially identical and partially different, when is the similarity between them great enough to allow us to generalize the suffixes that are seen in one, but not in the other, to both of them? This remains an unsolved problem.

7.7 From analogy to algorithm

How does one actually *find* analogies along the lines of *book : books :: dog : dogs* in a language? It turns out that questions of this sort are not at all easy to answer, and a large part of the work devoted to the *Linguistica* project has been aimed at providing answers to this question. In this section, I will describe two problems that seem simple enough, and are certainly typical, and try to give a sense of why they are not as simple as one might expect them to be. The first example is the treatment of gender and plural marking of adjectives in French; the treatment of parallel forms in a number of other languages, such as Spanish, would be similar. The second is the treatment of morphological patterns in a rich system like that of the Swahili verb. “Treatment” in this context means the breaking up of the string into substrings corresponding to morphemes and the correct formulation of a finite-state automaton (or its equivalent) to generate the observed patterns. Thus we address both the first and the second question articulated in the first section of this paper.

As I noted above, some pre-generative linguists took such questions very seriously—notably, Zellig Harris (1955, 1967) did (but see Hafer and Weiss 1974). Harris apparently believed that he had solved the problem through the computation of what he called successor frequency (and predecessor frequency) in a large corpus. By successor frequency, Harris meant a characteristic of a specific string, in the context of a specific corpus: given a string S of length n (typically the first n letters of a word), one considers only the subset of words in the corpus that begin with the string S (computer scientists would say: consider the set of words with the prefix S —but then computer scientists use the term *prefix* rather differently than linguists), and then one asks: in this subset, how many different letters are there in the $(n + 1)^{st}$ position (which is the position right after the string S)? That value is the successor frequency of string S , in the corpus.

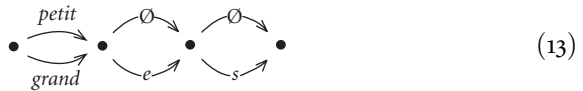
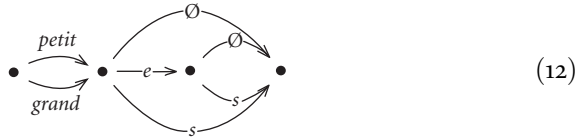
Harris believed that by calculating the successor frequency and the predecessor frequency at each point in each word of a corpus, he could find the morpheme boundaries (although Hafer and Weiss note that on the basis of their experiments neither choosing a threshold nor looking for a local maximum of successor frequency works very well in English). To make a long story short (see Goldsmith (2001, 2006) for the long version), such a purely local method does not work, and some more global characteristics of the overall grammar need to be taken into consideration, as we have already suggested.

Still, Harris's notion of successor frequency can serve as a useful heuristic for locating potential breaks, as the simple data in (1) suggest: the presence of the words *jump*, *jumped*, and *jumping* in a corpus leads to a successor frequency of three after the stem *jump*, just as it is after *walk*.

But successor frequency fails to work, even as a heuristic, when we turn to languages with much richer morphologies (that is, where the average number of morphemes per word is considerably higher than it is in English), and as linguists know, the morphological richness of English is on the poor side, as languages go.

The first case we will consider is that of the regular inflectional pattern of written modern French, which represents an earlier form of spoken French (some of this material is discussed in greater detail in Goldsmith and Hu (2004)). In the treatment of a subcorpus like *petit*, *petits*, *petite*, *petites*, *grand*, *grands*, *grande*, *grandes* (the masc. sg., masc. pl., fem. sg., and fem. pl. forms for *small*, *large*), the system we have described in Goldsmith (2006) will generate an FSA as in (11), and an algorithm described in Goldsmith and Hu (2004) generates the FSA in (12) rather than (13), which is the correct structure. The FSA in (11) misanalyzes the segmentation of the feminine plural forms, and (12)

correctly segments, but does not represent the correct grammar, which is that given in (13). In terms of analogy, all three systems capture the analogy *petit* : *petits* : *petite* : *petites* :: *grand* : *grands* : *grande* : *grandes*, but only (13) expresses the analogy *petit* : *petits* :: *petite* : *petites* and also *petit* : *petite* :: *petits* : *petites*. (In fact, it appears to me easier to understand the nature of the generalization being captured by looking at the FSA than by using the traditional notation associated with analogy expressed with colons.)



The two big questions are: does a natural complexity measure unambiguously choose (13) over (11) and (12), and do we have a good search procedure that finds (13)? A relatively brief summary provides a positive answer to the first question; the second is more difficult to answer, and I will leave it open for now. The complexity of an FSA is almost exactly equal to the sum of the informational complexity associated with each of its nodes plus that of each of its edges plus that associated with the labels on the edges. As noted above, the informational complexity is in each case the inverse log probability of the item in question. In (11), there are three nodes, each of which has roughly the same informational complexity, equal in this case to $\sigma = -\log S$, where S is the frequency of words that is described by this FSA in the corpus (that is, the total count of the words in this FSA divided by the total number of words in the corpus). The information complexity of the labels on each edge are also equal to the inverse log frequency of their usage, and *es* is a relatively rare suffix in French (i.e., there are relatively few feminine plural adjectives), and hence its informational cost is quite large. In addition, one must pay twice for the two pointers to each of the suffixes \emptyset and *s*, and there is one more node in (12) than in (11). Hence (12) turns out to be more costly than (11). By contrast, (13) is less

complex than either (11) or (12), despite the fact that it has one more node than (11). By avoiding positing a morpheme *es* (expensive because rare—it costs less than (11)), while by positing *s* only once, it costs less than (12).

I think this example clearly illustrates the basic point of this paper: formal complexity can, in many cases, be used to evaluate and compare alternative analyses, and algorithmic and information-theoretic complexity suffices to define the relevant complexity.

The second example we will look at represents still uncharted waters. It come from Swahili; consider (14), which gives a sample of some of the richness of the Swahili finite verb; I use the traditional Bantu terminology where appropriate. The positions indicated in this diagram illustrate subject markers, tense markers, object markers, verb roots, the passive/active marker, and the final vowel, respectively; there are also other affixes, such as a relative clause marker that can appear after the tense markers, which are not indicated here. There is little question but that the correct solution is formally much simpler than any of the partial solutions; algorithmic complexity will correctly identify an FSA as in (14) as a very simple grammar.

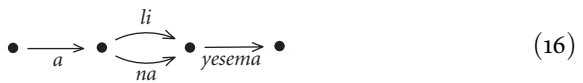


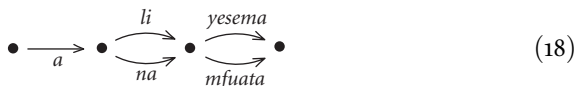
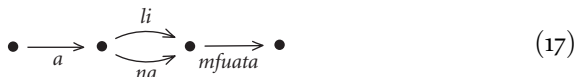
In order to even have a chance to discover these morphemes and the structure that lies behind them, we need to implement the notion of analogy in a richer fashion; what follows is taken from Hu *et al.* (2005).

We first look for elementary alignments between pairs of strings, as in (15), where m_1 or m_4 can be null, and m_2 or m_3 can be null. These elementary alignments can be found using the well-known string edit distance algorithm.



We expand these structures by finding ways to collapse them, either as suggested by (16), or as in (17) and (18).





But establishing a clear and workable algorithm to correctly collapse these FSAs is no simple task, in the presence of only a realistic amount of data (and it is not clear that increasing the amount of data available would change the difficulty in an essential way). The simple cases illustrated here work fine to collapse small FSAs when the difference between them is small. But the problem becomes harder quite quickly when we try to induce the correct structure, for example, of what is perhaps the structure best represented in the data, that found in the first two “columns” of (14), representing the subject markers and the tense markers. Because each column has a large number of possible morphemes in it, the subgeneralizations that we easily find—typified by the one in (18), which has a single subject marker (*a*) followed by two tense markers (*li* and *na*)—become harder and harder to analogize to.

Let’s be a bit more specific, to make concrete what we’re talking about. In a corpus of 25,000 Swahili words (4,100 distinct words among them), we find 254 three-state FSAs with the methods we have sketched, and of these, virtually all of them are linguistically reasonable; the place where the strings are cut are, indeed, morpheme boundaries from the linguist’s perspective. These three-state FSAs (and I have sketched the top eight in (19–26)) can be ranked with respect to how much information they compress: those that compress a good deal of information are necessarily those that express a large number of words with relatively few edges in the automaton. In theory, that kind of compression can happen in either of two ways: by specifying an FSA with a single stem but a wide range of affixes, or by specifying an FSA with a smaller set of affixes and a wide range of stems. It turns out that the latter is by far the most common kind of generalization obtained.

The task now is to generalize, which is effectively just another way of saying to learn what the morphological pattern of Swahili is. As far as I can see, there is little or nothing that we can posit as a simple innate premise that will help, nor will appealing to analogy help us, because the question now is really: when should two (or more) patterns be treated as analogous? Now, it is very likely true that if these strings of letters were labeled as the morphemes that

they are (that is, if the labels told us more than just the phonemes: if they furthermore identified the functional category of the morpheme), our task would be considerably lightened. But taking that information for granted seems to me like question-begging. Swahili, just like most languages, often employs the same sequence of phonemes to realize different morphemes (for example, the subject and object markers for various person and number classes is the same: *tu* marks both subject and object marker for first-person plural, etc.) It is morphological analysis, and the inference of a morphological generator, that is an important step on the way to understanding the morphological identity of strings of letters (or phonemes); we risk circularity if we assume that knowledge of morpheme identity can serve as the basis of our knowledge of the morphological grammar. We would like to understand how a learner would generalize by recognizing the identity of the prefix *a* in patterns (19), (20), (22), (24);¹⁶ but *a* is the most common phoneme and also the most common morpheme in the language, and occurs with several functions; mere phonological identity is simply *not* enough to lead the learner to treat all occurrences of *a* in the same way.

$$\left\{ \begin{array}{l} a \text{ Subject} \\ wa \text{ Markers} \end{array} \right\} \left\{ \begin{array}{l} 55 \text{ stems:} \\ baki \\ ende \\ fanye \\ \dots \end{array} \right\} \quad (19)$$

$$\left\{ \begin{array}{l} a \text{ Subject} \\ m \text{ Markers} \end{array} \right\} \left\{ \begin{array}{l} 17 \text{ stems:} \\ cheni \\ kaanguka \\ kapoteza \\ \dots \end{array} \right\} \quad (20)$$

$$\left\{ \begin{array}{l} 17 \text{ stems:} \\ akaongez \\ alifany \\ ameja \\ \dots \end{array} \right\} \left\{ \begin{array}{ll} NULL & \text{active} \\ w & \text{passive} \end{array} \right\} \quad (21)$$

¹⁶ And perhaps (25): the system posits *ana* as a prefix, and it is an inductive leap to treat this as the concatenation of *a* and *na* at this point.

$$\left\{ \begin{array}{ll} a & \text{Subject} \\ wa & \text{Markers} \end{array} \right\} \left\{ \begin{array}{l} 14 \text{ stems:} \\ \textit{changa} \\ \textit{heshimuni} \\ \textit{lilokataa} \\ \dots \end{array} \right\} \quad (22)$$

$$\left\{ \begin{array}{l} 12 \text{ stems:} \\ \textit{akawaachi} \\ \textit{amewaweke} \\ \textit{fany} \\ \dots \end{array} \right\} \left\{ \begin{array}{ll} a & \text{default ending} \\ \textit{eni} & \text{plural imperative} \end{array} \right\} \quad (23)$$

$$\left\{ \begin{array}{ll} a & \text{Subject} \\ & \text{Marker} \end{array} \right\} \left\{ \begin{array}{ll} \textit{li} & \text{Tense} \\ \textit{na} & \text{Markers} \end{array} \right\} \left\{ \begin{array}{l} 11 \text{ stems:} \\ \textit{batiza} \\ \textit{chaguliwa} \\ \textit{kwenda} \\ \dots \end{array} \right\} \quad (24)$$

$$\left\{ \begin{array}{ll} \textit{ana} & \text{Subject Marker} \\ & \text{and Tense Marker} \end{array} \right\} \left\{ \begin{array}{ll} \textit{NULL} & \text{default} \\ \textit{ye} & \text{Rel Clause marker} \end{array} \right\} \\ \times \left\{ \begin{array}{l} 10 \text{ stems:} \\ \textit{fanana} \\ \textit{ishi} \\ \textit{kuja} \\ \dots \end{array} \right\} \quad (25)$$

$$\left\{ \begin{array}{l} 18 \text{ stems:} \\ \textit{akili} \\ \textit{bahari} \\ \textit{dunia} \\ \dots \end{array} \right\} \left\{ \begin{array}{ll} \textit{NULL} & \text{default} \\ \textit{ni} & \text{postposition} \end{array} \right\} \quad (26)$$

We are currently working on a method to link the low-level FSAs illustrated in (19–26) to the larger, simpler, and correct pattern, that of (15), and I will sketch the intuition that lies behind it. These FSAs can be thought of themselves as expressions (for example, by alphabetizing all the elements in a column and concatenating them with a punctuation marker between them), and we can establish a distance measure across pairs of string expressions which we can then use to hypothesize which items should be collapsed to

form a larger generalization. When two or more morphemes—especially high-frequency morphemes—appear in the same column (that is, in a paradigmatic morphological relationship), then they may be analyzed as likely alternatives for the same morphological position.

This is easier to explain with a real example. There are several high-frequency FSAs that begin with the subject marker *a*, followed by two alternative tense markers, followed by a set of verbal stems. In the first case, the two tense markers are *li* and *na*; in the second, the two tense markers are *li* and *me*; in the third, they are *ka* and *na*; in the fourth, *ka* and *li* (I have not listed these FSAs here). We can capitalize upon each of these pairings to create a distance metric among these morphemes with this information, increasing the simplicity of assigning them to the same morphological position. We do this in order to overcome the problem of the sparsity of the data: we never find a single stem in a finite corpus appearing in all of its possible forms; what we need to do is use the partial information that the data actually provide, and much of that information is bundled into the observation that various subsets of morphemes appear in the same position of the word—and we can infer that even before we have a clear global understanding of what the overall structure of the word is. In a sense, that's the key to understanding learning: understanding how we can incrementally advance the analysis of the data, through analyzing the data, even though we have not yet achieved a global understanding of how everything fits together. In this case, the appearance of a pair of stems (*keti*, *mtuma*) appearing with the subject marker *a* and three of the four tense markers (*ki*, *li*, *na*, in fact) strongly supports the hypothesis that they are all realizations of the same morphological position. The sense in which this is true can be mathematically formulated and integrated into the search algorithm. But considerable work remains if we are to correctly induce the simple, and globally coherent, morphological structure of forms like the Swahili verb.

7.8 Discussion and conclusion

We have covered—or at least touched on—quite a number of topics, all closely joined by the question of how morphology can be learned. We have focused on the task of learning to segment words into morphs and discovering the grammar which puts them back together. This task is already difficult enough, but I hope it is clear that in a sense this task is a surrogate for the larger and more difficult task of segmenting entire utterances (into the pieces we call words) and discovering the grammar which puts them back together.

In the case of morphology, there is little or no hope that an appeal to a magical slate of innate principles will greatly simplify the task (I refer, of course, to an information-rich Universal Grammar). As far as learning morphology is concerned, Locke was surely right: all the reasoning is search and casting about; it requires pains and application. But we must not lose sight of the fact that even if language learning means searching and casting about on the part of the learner, there still must be an overarching model which describes what it is that is being sought. It seems to me that only a highly mathematical model which comes to grips with the complexity (in the technical sense) of the hypothesis has even a chance of shedding light on the problem of language learning. And if this conjecture is correct, then it seems to me almost a certainty that the same learning mechanisms can be used to induce a syntax as well. While it is not logically impossible that learning morphology requires a rich and powerful learning theory and learning syntax does not, such a state of affairs is highly unlikely at best.

A word, in closing, is perhaps appropriate regarding the relationship between the kind of linguistic work we have sketched and the study of child language acquisition, since it is only natural to ask what connection is being posited between the two. The two answer different questions: the linguist asks how language *can be* learned; the psycholinguist asks how language *is* learned. Each has his work cut out for him. If the linguist had several adequate theories of how language could be learned, the psycholinguist could figure out which was the right one—but the linguist does not. If the psycholinguist could provide an account of how language *is* learned, we would have at least one answer to the question as to how language can be learned—but the psycholinguist does not. We are making progress, I think, regarding the models on the market for morphology learning, and some aspects of phonology learning, and there is a time-honored law according to which once we find *one* way to accomplish something, several more will present themselves virtually overnight.

These questions are reflections of an old and traditional debate between rationalist and empiricist inclinations in the study of mind, but the most familiar versions of how *both* schools have treated language acquisition are, in my view, coarse oversimplifications. Rationalists of the principles-and-parameters sort attempt to account for language learning by denying its existence, and hoping that the variation across the world's languages will simply go away, while empiricists of the old school hope that knowledge can be reduced to memory. Both of these are losing strategies, in my view, and I have tried to offer some specifics with regard to one small, but not insignificant, part of language learning. It is an empiricist account that sets a high bar for formal grammatical accounts of the relevant data.

7.9 Appendix

Let us consider how the probability of a corpus changes when we begin our word discovery process. Originally, the lexicon consists of the observed letters in the corpus. Our first guess will add the string TH to the lexicon. When we add the element TH, the log probability of the corpus is changed in three ways. First, the total number of words in the corpus decreases by the number of THs found in the corpus (that may not be obvious, but it is true, if you think about it). Second, the total number of Ts and Hs also decrease (since a T that is followed by an H is no longer parsed as a T, but rather as part of a TH), and hence the probability of both Ts and Hs decreases, since those probabilities are based on observed frequencies. (Note, by the way, that this illustrates the point that even frequencies are theory-dependent notions!) Third, the probability of the substring TH has gone up considerably, because it had previously been calculated as the product of the probabilities of T and H independently, but now it is calculated on the basis of the observed frequency of the sequence TH. The actual change in log frequency is $-N\Delta N + [t]\Delta[t] + [h]\Delta[h] + [th] \log \frac{\text{freq}_2(th)}{\text{freq}_2(t)\text{freq}_2(h)}$, where N is the original length of the corpus and thus the number of words on the first analysis, ΔN is the log ratio of the count of words *after* versus *before*, i.e., $\log \frac{N - \text{number of THs}}{\text{number of letters}}$, $[t]$ and $[h]$ are the number of *T*s and *H*s in the original corpus, $\Delta[t]$ is the log ratio of the counts of *T* after vs before and likewise for $\Delta[h]$, and $[th]$ is the number of substrings *TH* found in the corpus; $\text{freq}_2(x)$ is the frequency of x in the second model, that in which TH is interpreted as a single lexical item. Note that ΔN , $\Delta[t]$, and $\Delta[h]$ are all negative.

Expanding Analogical Modeling into a general theory of language prediction

Royal Skousen

8.1 The core theory

Analogical Modeling (AM) is a general theory for predicting behavior. It can also be considered a system of classification or categorization according to a particular set of outcomes. Predictions are directly based on a dataset of exemplars. These exemplars give the outcome for various configurations of variables, which may be structured in different ways (such as strings or trees). A common method in AM is to define the variables so that there is no inherent structure or relationships between the variables (that is, each variable is defined independently of all the other variables). In this case, the variables can be considered a vector of features. In the dataset, each feature vector is assigned an outcome vector. The dataset is used to predict the outcome vector for a test set of various feature vectors for which no outcome vector has been assigned (or if one has been assigned, it is ignored).

In AM we distinguish between the core theory and its application to language. In terms of the theory, the goal is to predict the **outcome** for a set of conditions referred to as the **given context** (sometimes the given context is referred to as the **test item**). From the given context, we construct more general versions of that context, which we refer to as **supracontexts**. Our goal is to predict the behavior (or outcome) of the given context in terms of the behavior of its supracontexts. The source for determining those behaviors comes from a **dataset of exemplars**; for each exemplar in the dataset, the outcome is specified. These exemplars, with their own specifications and associated outcomes of behavior, are assigned to the various supracontexts defined by the given context. Supracontexts that behave uniformly (referred

to as **homogeneous** supracontexts) are accepted, with the result that exemplars contained within the homogeneous supracontexts can be analogically used to predict the behavior of the given context. The exemplars found in nonuniformly behaving supracontexts (referred to as **heterogeneous** supracontexts) cannot be used to make the analogical prediction for the given context. The term *nonuniformity* means that a heterogeneous supracontext has a plurality of subcontexts and a plurality of outcomes (that is, exemplars within the supracontext not only have different outcomes but they are also found in different subspaces of the contextual space). Finally, the relative probability of using a homogeneous supracontext is proportional to the square of its frequency, while the probability of using a heterogeneous supracontext is zero. (For a basic introduction to AM and how it works, see Skousen, Lonsdale, and Parkinson 2002: 12–22 or Skousen 2003.)

AM differs considerably from traditional analogical approaches to language. First of all, traditional analogy is not explicit. In the traditional practice of analogy, virtually any item can serve as the exemplar for predicting behavior, although in practice the first attempt is to look to nearest neighbors for the preferred analogical source. But if proximity fails, one can almost always find some item considerably different from the given item that can be used to analogically predict the desired outcome. In other words, if needed, virtually any occurrence with a minimum of similarity can serve as the analogical source. AM, on the other hand, will allow occurrences further away from the given context to be used as the exemplar, but not just any occurrence. Instead, the occurrence must be in a homogeneous supracontext. The analogical source does not have to be a near neighbor. The probability of an occurrence further away acting as the analogical model is usually less than that of a closer occurrence (all other things being equal), but this probability is never zero (providing the occurrence is in a homogeneous supracontext). Proximity is important in AM, but it is not the only factor.

A second important property of AM is that analogy is not used as a stop-gap measure to be used whenever the rules fail to account for the behavior. Instead, everything in AM is analogical. Rule-governed behavior, so called, comes from homogeneous groups of occurrences that behave alike, leading to gang effects that enhance the probability of using occurrences in frequently occurring homogeneous supracontexts. In other words, categorical and regular/exceptional behaviors are accounted for in terms of exemplars, not categorical rules or regular rules with lists of exceptions.

Another important property of AM is that it does not determine in advance which variables are significant and the degree to which these variables determine the outcome (either alone or in various combinations). In addition, AM

does not have a training stage except in the sense that one must obtain a database of occurrences. Predictions are made “on the fly”, and all variables are considered equal a priori (with certain limitations due to restrictions on short-term memory). The significance of a variable is determined locally – that is, only with regard to the given context. Gang effects are related to the location of the given context and the amount of resulting homogeneity within the surrounding contextual space.

One simplified way to look at AM is in terms of traditional rules, where the term *rule* basically stands for the supracontext and its associated behavior. In trying to predict the behavior of the given context, we consider all the possible rules that could apply. We eliminate those rules that behave nonuniformly (that is, the rules with heterogeneous supracontexts). All uniformly behaving rules (the rules with homogeneous supracontexts) are then applied, with the probability of applying a given homogeneous rule proportional to the square of its frequency. One important aspect of AM is that each rule’s homogeneity can be determined independently of every other rule. This property of independent determination of uniformity means that we can examine a rule’s uniformity without having to determine whether any subrule (that is, any more specific version of the rule) behaves differently.

AM is computationally intensive. For each variable added to the specification of a given context, both the memory requirements and the running time doubles (so if there are n variables in the given context, the memory and time are of the order 2^n). This problem of exponential explosion has been theoretically solved by redefining AM in terms of Quantum Analogical Modeling (QAM), a quantum mechanical approach to doing AM. The main difference is that everything is done simultaneously in QAM, in distinction to the sequential application that AM is forced to follow. Still, the same basic procedure is followed, only the system of rules (or supracontexts) is now treated as a quantum mechanical one:

- (1) all possible rules for a given context exist in a superposition; the initial amplitude for each rule is zero;
- (2) the exemplars are individually but simultaneously assigned to every applicable rule; after all the exemplars have been assigned, the resulting amplitude for each rule is proportional to its frequency (that is, to the number of exemplars assigned to that rule);
- (3) the system evolves so that the amplitude of every heterogeneous rule becomes zero, while the amplitude of each homogeneous rule remains proportional to its frequency (that is, to the number of exemplars originally assigned to that rule);

- (4) measurement or observation reduces the superposition to a single rule where the probability of it being selected is proportional to its amplitude squared.

See Skousen 2002: 319–46 for an introductory essay on treating AM as a quantum mechanical system. For a complete discussion of how QAM works, see Skousen 2005.

One notices here that nothing in the core of AM specifies how AM is to be applied to language. All such language applications have their own linguistic assumptions, and it is an open question not directly related to AM itself concerning what those assumptions should be. But by choosing various assumptions and seeing what kinds of predictions AM makes about the behavior, then by comparing the predicted behavior to the actual behavior, we can assess the empirical validity of those linguistic assumptions.

A similar situation exists in quantum mechanics, which seems appropriate to bring up here since QAM itself is a quantum mechanical system. As explained by Charles Bennett, there is a “set of laws” (like the Ten Commandments, as he puts it) that form the basics of quantum mechanics (QM), but QM has to be applied in order to serve as a theory of physics: “For most of the 20th century, physicists and chemists have used quantum mechanics to build an edifice of quantitative explanation and prediction covering almost all features of our everyday world.” The core theory is actually very simple, but the resulting edifice is complex and evolves. Yet in all instances, QM involves applying the core theory and making hypotheses regarding the underlying physical system. If the resulting application of the theory works, we accept the hypotheses as representing, in some sense, physical reality. For a pictorial representation of this point, see Bennett 1999: 177–80.

AM is a general theory of predicting classification and is not restricted to linguistic problems *per se*. For instance, one can use AM to predict various kinds of nonlinguistic outcomes, such as determining whether different mushrooms are poisonous or not, providing medical diagnoses based on symptoms and lab tests, and predicting party affiliation on the basis of voting patterns (for various examples, see Lonsdale 2002).

8.2 Basic structural types

The current AM computer program treats the n variables defined by a given context as n independent variables, which means that any linguistic dependencies between the variables must be built into the variable specifications. Very seldom can we construct cases where there are no linguistic dependencies in

the variable specification (although there will always be behavioral dependencies between the variables). One possible example of linguistic independence of variables involves the social features for specifying terms of address in Arabic (see Skousen 1989: 97–100), which has social variables like age of speaker, gender of speaker, and social class relationship. Very often the linguistic task involves strings (in phonology) or trees (in morphology and syntax). The general approach in Skousen 1989 was to treat strings of characters as variables for which the position was specified. For instance, in predicting the spelling of the initial /h/ sound in English words (as either *h*, *wh*, or *j*), positional aspects were included in the definition of each variable, such as “the first vowel phoneme” and “the phoneme that *immediately precedes* the third vowel.” This kind of variable specification allows the AM computer program to make the analysis, but it is not realistic since it requires that everything be lined up in advance so that the strings can be compared. Cases of metathesis or identical syllables in different positions are ignored, nor can they be readily handled in such a restricted version of AM.

These kind of specifications have led to the use of zeros for variables. For instance, if a word has only one syllable, then the nonexistent second and third syllables are marked with zeros. But then there is the question of how to specify such nonexistent syllables. If we mark the nuclear vowel for such a syllable with a zero, do we also mark the syllable’s onset and coda with zeros, even though those zeros are redundant? One possibility is to refer to such predictable zeros as redundant variables and to ignore them when making analogical predictions (the general way of proceeding in Skousen 1989). If all zeros (both essential and redundant) are counted, then there is the possibility that the analogical prediction will be overwhelmed by excessively specified zeros. But there is also the possibility that we may want to count all the zeros (for some discussion of this issue, see Skousen, Lonsdale, and Parkinson 2002: 40–2). The important point here is that the problem of the zeros results from trying to account for strings as if they were composed of unordered symbols. One major issue that linguistic applications of AM must deal with then is how to treat strings and trees as they actually are rather than trying to define them as sets of independent variables. In the remainder of this section, I outline several different approaches to structures that one would want to use in linguistic analyses. It should be pointed out, however, that the current AM computer program has not yet been revised to handle these structures directly.

8.2.1 *Strings of characters*

A more reasonable approach for a string of characters would be to allow any possible sequence of substrings of a given string to count as a supracontext.

For instance, if the given string is *abc*, the supracontexts would include examples like **abc*, *a*b*c*, **ab*c**, *a*b*, and **c*, where the asterisk stands for any string, including the null string. Thus the supracontext **abc* would include any string ending in *abc* while **ab*c** would include any string containing *ab* followed by *c*. The most general supracontext would be simply *** (that is, this supracontext would contain all possible strings, including the null string). For *n* characters in a given string, there will be a total of $2 \cdot 3^n$ supracontexts for which we will need to determine the heterogeneity, but (as already noted) this can be done independently for each supracontext by determining its heterogeneity with respect to the outcomes and the subcontexts for the data items assigned to that supracontext. The total number of supracontexts for this kind of string analysis is also exponential (like the 2^n for when the *n* characters are all independent variables), but that number ($2 \cdot 3^n$) increases at a greater exponential rate.

8.2.2 *Scalar variables*

The AM computer program assumes that the variables specified by the given context are categorical and discrete. The question then arises of how to deal with scalar variables, ones that represent degrees of a property. Scalars can be mathematically treated as real numbers, but this leads to extraordinary problems with the number of possible supracontexts since theoretically every possible real number interval could count as a supracontext, which ends up defining a nondenumerably infinite set of supracontexts. I would propose, instead, that continuous scalars be analyzed as a sequence of finite intervals (that is, we will quantize the scalar). Having made that decision, we can then decide how to determine the supracontexts for a given sequence of finite intervals.

As an example, consider how we might apply this quantization to the problem of voicing onset time and the ability of speakers to predict whether a given stop is voiced or voiceless. Our task is to model how speakers interpret artificial stops with varying lengths of nonvoicing after the release of the stop. In this case, the data comes from experiments testing the ability to distinguish between /b/ and /p/ in English. The variables center around the problem of dealing with a time continuum. In applying AM to this problem, I assume that time should not be treated as a real number line. Instead, time will be broken up into a sequence of finite intervals of time, all equal in length, as described in Skousen 1989: 71–6. Given an overall length of about 50 msec between instances of /b/ and /p/, let us break up this overall length into five intervals of 10 msec each, so that instances of voiced stops are represented as

xxxxx and voiceless stops as *ooooo* (where *x* stands for voicing and *o* for nonvoicing). For simplicity of calculation, I will assume that there is in the dataset but one occurrence of each stop, /b/ and /p/. The question then becomes: What are the supracontexts for the intermediate but nonoccurring given contexts (namely, *oxxxx*, *ooxxx*, *ooox*, *oooox*)? I will here consider three possibilities for the supracontexts defined by the particular given context *ooxxx*:

- (1) We treat each single continuous sequence of intervals as a possible supracontext. The number of homogeneous supracontexts, in this case, will be quadratic – namely, $n(n+1)/2$:

<i>given context</i>	/b/ outcome	/p/ outcome	probability
xxxxx	15	0	1.000
oxxxx	10	1	0.909
ooxxx	6	3	0.667
ooox	3	6	0.333
oooox	1	10	0.091
ooooo	0	15	0.000

- (2) We treat each *o* and *x* and its position as an independent variable (this is how the problem is treated in Skousen 1989: 71–6). This means that any subset of the five variables will define the possible supracontexts. The number of homogeneous supracontexts will be exponential (to the scale of $2^n - 1$), which means that in comparison with the previous case, the shift in predictability will be sharper. We get the following predicted chances for /p/ and /b/:

<i>given context</i>	/b/ outcome	/p/ outcome	probability
xxxxx	31	0	1.000
oxxxx	15	1	0.938
ooxxx	7	3	0.700
ooox	3	7	0.300
oooox	1	15	0.063
ooooo	0	31	0.000

- (3) Finally, we treat the sequence of intervals as a string and permit any set of nonoverlapping substrings to serve as a distinct supracontext. In this case, the shift in predictability will be sharper than in the second case (but also exponential) since the number of homogeneous supracontexts will have the exponential factor $2(3^n - 1)$ rather than the $2^n - 1$ of the second case:

<i>given context</i>	<i>/b/ outcome</i>	<i>/p/ outcome</i>	<i>probability</i>
XXXXX	484	0	1.000
OXXXX	160	4	0.976
OOXXX	52	16	0.765
OOOXX	16	52	0.333
OOOOX	4	160	0.024
OOOOO	0	484	0.000

For each of these three cases, we can determine which interval length allows for the best fit for the actual experimental results for predicting /b/ versus /p/ (see Lisker and Abramson 1970, cited in Skousen 1989). For an overall interval of 50 msec, we get the following:

<i>type of analysis</i>	<i>number of homogenous supracontexts</i>	<i>number of intervals</i>	<i>length of interval</i>
(1) a single continuous substring	$n(n+1)/2$	10	5 msec
(2) n independent variables	$2^n - 1$	7	7 msec
(3) any nonoverlapping sequence of substrings	$2(3^n - 1)$	5	10 msec

Lehiste 1970 (cited in Skousen 1989) provides evidence that speakers can distinguish between sound durations differing as little as 10 milliseconds, which means that the last case (which defines the supracontexts as any nonoverlapping sequence of finite intervals) is the one that best corresponds with experimental results for humans trying to distinguish between artificial versions of /b/ and /p/ in terms of voicing onset time.

8.2.3 *Unordered hierarchical structures (branching hierarchical sets)*

Unordered hierarchical structures are found in semantics. The supracontexts for a given hierarchical set are subsets that generalize by moving up the hierarchy, thus accounting for hyponymy. Semantic variables (or features) defined for lower, more specific subsets may be ignored in higher, more general subsets.

The need for localized restrictions on the use of semantic features is well exemplified in an attempt to analyze and predict the behavior of Chinese classifiers, found in some unpublished work by my colleague Dana Bourgerie, presented at the 2000 Analogical Modeling conference at Brigham Young University ("An Analysis of Chinese Classifiers: Issues in Dealing with Semantic Variables in the AML Framework"). Some of the classifiers examined by Bourgerie were:

<i>gè</i>	general classifier for people, round things, and things of indeterminate form
<i>zhāng</i>	for open flat things (maps, tables, tickets, etc.)
<i>tiáo</i>	for long, thin things (fish, leg, boat, cucumber, a long bench, etc.)
<i>zhī</i>	for long, branch-like things (e.g. pen, gun, candle, etc.)
<i>bǎ</i>	things with handles (e.g. umbrellas, swords, etc.)
<i>jiān</i>	mostly for rooms
<i>běn</i>	for bound things such as books

The need for an analogical model results from a great deal of variability in actual usage between speakers in selecting the appropriate classifier as well as the extension of classifiers to new objects.

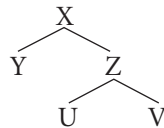
Bourgerie's variables were as follows: relative size (*s*, *m*, *l*), flat (+, -), long (+, -), narrow (+, -), three-dimensional (+, -), handle (+, -); every instance of usage involving a classifier in his dataset was defined in terms of these specific variables. Given what we know now, the size variable should have been converted to a discrete scalar (something like --, -+, and ++ to stand for small, medium, and long, respectively); I will make that conversion here to simplify the description. In other words, the semantic description of every noun in the dataset can be analyzed as a sequence of pluses and minuses. In Bourgerie's preliminary work, pluses and minuses were assigned in all cases. For example, *bǎ* is expected for objects with a handle, even though other items with handles, such as a gun, take *zhī*. On the other hand, some objects, such as a boat or long bench (which take the classifier *tiáo*), do not ordinarily have handles, yet '-handle' was assigned to this classifier. And finally for some nouns referring to people (which take the classifier *gè*), handles would seem implausible, although one could imagine it! Implausible or not, '-handle' was assigned to words taking the most general classifier. And similar overloading of minus-valued variables occurred for other specific classifiers. The overall result was that the classifier *gè*, being the most general classifier, had more minuses for the words assigned to it – and especially more minuses than words assigned to the other (more specific) classifiers.

The problem with assigning '-handle' (and similarly for other specific variables) to all the nouns in the dataset is that when predicting the classifier for any given word, the minuses dominate, with the result that the general classifier *gè* consistently swamps the prediction, even when we are predicating an item close to words that take one of the more specific classifiers. To get the right results, we need to restrict '+/- handle' to smaller groups of words where they characteristically are found. So we may mark kitchen utensils as to whether or not they have a handle, but not the refrigerator, oven, sink, dishwasher, counters, tables (although they could have them). Where the handle helps to distinguish

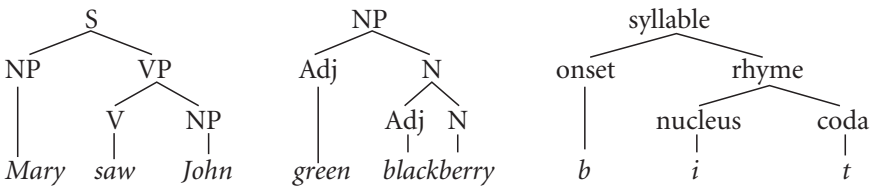
between closely associated objects that are named differently, the variable should be assigned, but otherwise not. A table could have a handle, but such a table doesn't have a different name, so we do not specify it as a variable in such cases. The vast majority of words could always be marked as '–handle' (such as a cloud, a tree, a lake, a newspaper, philosophy, war, etc.), but AM shows that we cannot semantically analyze every word as plus or minus for every possible semantic variable. This may seem obvious: Would we really want to mark virtually every object in the world as '–human'? Semantic variables are defined within only certain restricted domains. In applying AM to Chinese classifiers, Bourgerie marked every word in the dataset as either '+handle' or '–handle' and soon discovered that such a decision clearly made the wrong predictions.

8.2.4 Ordered hierarchical structures (trees)

Ordered hierarchical structures obviously have both order and hierarchy and are commonly referred to as trees. Given a particular tree as a given context, the supracontexts are defined as subtrees of the given tree. For instance, our given contexts and the data items may best be represented as trees and we may wish to predict some behavior given such a tree. The following simple right-branching structure is of interest in many different situations:



We find uses of it in specifying syntax, morphology, and syllable structure:



In attempting to predict some outcome based on the pronunciation for the last item, *beet* /bit/, we could restrict the supracontexts for the given context (namely, the tree itself) to combinations of categories that occur only at the same level in the tree; for instance, we could examine all syllables with the same onset, or with the same rhyme (nucleus and coda), or with the same nucleus, or with the same coda – but not with the same onset and the same nucleus or with the same onset and the same coda since those categorical combinations do not occur at the same level in the tree. Of course, we would

only want to do this if there was evidence that such a restriction on supra-contextual construction would predict language behavior. In other words, the decision is more an empirical one than one that impinges on the question of whether AM is a correct theory.

We get similar hierarchical problems in specifying distinctive features. As discussed in Skousen 1989: 53–4, we cannot treat distinctive features as if they are independent variables. Suppose we compare *beet* /bit/ with two possible words, each of which differs from /bit/ in three distinctive features. If we treat this problem as a set of twelve variables, the distance between *beet* and *bought* is the same as between *beet* and *mid*:

(a) *three-feature difference restricted to one phoneme:*

	consonant	vowel	consonant
/bit/	oral stop labial voiced	spread high front tense	oral stop alveolar voiceless
/bɔt/	oral stop labial voiced	round low back tense	oral stop alveolar voiceless

(b) *three-feature difference spread across three phonemes:*

	consonant	vowel	consonant
/bit/	oral stop labial voiced	spread high front tense	oral stop alveolar voiceless
/mɪd/	nasal stop labial voiced	spread high front lax	oral stop alveolar voiced

Yet experimental evidence from perceptual studies show that speakers perceive *beet* and *bought* as phonetically close, while *beet* and *mid* are not especially close (see Derwing and Nearey 1986, cited in Skousen 1989). If we treat distinctive features as independent variables, we incorrectly predict an equality of phonetic similarity for this example. One way to correct this would be to define the given contexts in terms of phonemes and basic syllable structure, which would mean that there is only one difference between *beet* and *bought*, but three between *beet* and *mid* (this is how it is done in Skousen 1989). But another possibility would be to define distinctive features for only phonemic nodes within syllable tree structures, thus restricting feature similarity to apply only at isolated places in the tree.

8.3 Control over the analogical set

The general theory of analogical modeling (AM) allows for various ways of using the analogical set to predict outcomes (although the quantum version of it, QAM, does not). Here I review this aspect of AM.

8.3.1 *Reacting to a previous prediction*

One important point is to recognize that analogical modeling allows for the ability to reexamine a given analogical set or to redetermine it under various conditions. A speaker may, for instance, produce a particular outcome, but then not like the results and so produce a different outcome. The speaker does not get caught in an infinite loop, continually producing, say, the most favored outcome or randomly producing outcomes, thus leading to the repetition of the more frequent outcomes. Consider, for instance, the following two examples from my own children's speech (cited in Skousen 1989: 85–6):

Nathaniel (5 years, 10 months)

Looking at a picture of the Grand Canyon, Nathaniel keeps trying to produce the plural *cliffs*: /klɪ'ftəz/, /klɪfs/, /klɪvz/, /klɪfs/

Note that Nathaniel's sequence of productions is not constantly repetitive (as if it were /klɪfs/, /klɪfs/, /klɪfs/, /klɪfs/, ...).

Angela (6 years, 10 months)

The possessive form *Beth's* is pronounced first as /bes/ and then immediately followed by /be'θəz/.

Angela: How do you add the *s* to *Beth*? It's hard to say. How do you say it?

Royal: I say /bes/ [bes:].

Angela: I say /beθ/ like *Beth house* /beθ haus/.

Note that Angela produced a sequence of three different possibilities: /bes/, /be'θəz/, /beθ/.

Angela (7 years, 11 months)

The plural form *ghosts* is pronounced initially as /gousts/, then as /gous/, and is finally followed by the question "How do you pronounce that?"

Similarly, suppose we have a nonce word (written out) and ask someone to pronounce it; then no matter what they say, we say that it's wrong and ask for an alternative pronunciation. Our subjects do not go into an infinite loop; instead they will typically produce a sequence of different responses. An example is the nonce word YEAD, which might be pronounced alternatively as /yid/, /yed/, /yeid/.

For each new prediction, we could let the analogical set be redetermined from scratch but with all data items having the forbidden outcome eliminated so that those exemplars will not play a role in constructing the analogical set, especially since the original analogical set may provide only one possible outcome. Or

maybe one has a choice: Try the original analogical set first; if that fails, then revert to redetermining it by omitting the forbidden outcome.

8.3.2 *Random selection versus selection by plurality*

Another aspect dealing with control over the analogical set is the choice between random selection of an outcome and selection by plurality (discussed in Skousen 1989: 82–5). Psycholinguistic experiments show that speakers of all ages can reproduce probabilistic behavior by applying random selection to the analogical set. But as speakers grow older, by about age 8, they are also able to select the most frequent outcome, especially when they expect or want to make some gain from the choice of outcome. It can also be shown that if the choice involves some loss, then the most advantageous decision is to choose the least frequent outcome (discussed in Skousen 1992: 357–8). The ability to select by plurality would apparently require some kind of sampling or analysis of the analogical set, perhaps as it is being determined.

8.3.3 *Restricting morphological extension*

Another issue involving restrictions on the use of the analogical set asks whether there are any limits besides heterogeneity in preventing the overuse of analogy. Consider, for instance, the analogical prediction of the past tense in English for the verb *be*. The question here is whether the verb *see* (with its exceptional past-tense form *saw*) can be used as an exemplar in predicting the past-tense form for *be*:

/si/ : /sə/ :: /bi/ : /bə/ (that is, *see* : *saw* :: *be* : *baw*)

This analogical extension seems highly unlikely. One might argue that such an analogy is difficult simply because the chances of forgetting the past-tense *was/were* for the very frequent verb *be* are virtually negligible. But the question still remains: Is *baw* even possible? And if so, is there any way besides appealing to heterogeneity to restrict the applicability of *saw*? Here heterogeneity may not work since *see* is such a close neighbor to *be*, at least close enough to allow it to analogically apply to *be*.

Since we know the analogical set can be examined prior to using it, perhaps the speaker can reject an unrecognizable past-tense form. One could argue that the analogical set provides only results, not how those exemplars are derived. The analogical *baw* could therefore be possible, but at the same time unrecognizable, thus one could simply avoid using it. A similar case involves verbs of the form CX-Cɔt:

<i>alternation</i>	<i>example</i>	<i>extension</i>
ing-ɔt	bring-brought	sting-stought
ink-ɔt	think-thought	drink-drought
ach-ɔt	catch-caught	latch-lought
ai-ɔt	buy-bought	try-trought
ich-ɔt	teach-taught	reach-rought
ik-ɔt	seek-sought	tweak-twought

Is heterogeneity sufficient to prevent any of these analogies from applying? Probably not. But these analogies could nonetheless be rejected by speakers since the resulting past-tense forms are unrecoverable – that is, speakers are unable to determine what verb the past-tense form stands for. A past-tense prediction like *stought* would imply only that the analogical present-tense verb form began with *st*.

One could propose that unique alternations can never be extended analogically, but this is definitely false. We have, for instance, analogical extensions based on the noun *ox* and its uniquely exceptional plural form *oxen* (thus *axen* for the plural of *ax* and *uxen* for the nonce *ux*). But note that in these cases the singular forms *ax* and *ux* are recoverable from *axen* and *uxen*. The question may not be one of uniqueness, but rather recoverability.

8.4 Specifying the variables

One important aspect of AM is that we not restrict our analysis to just the important or crucial variables. We need to include “unimportant” variables in order to make our predictions robust. Consider, for example, the indefinite article *a/an* in English. Knowing that the following segment, whether consonant or vowel, “determines” the article (*a* for consonants, *an* for vowels), one could specify only the syllabicity of the following segment and thus predict *a/an* without error. Basically, we would be specifying a single rule analysis for the indefinite article. Yet in modeling the behavior of the indefinite article, AM specifies in addition the phonemic representation for that first segment in the following word as well as the phonemes and syllabicity for other segments in that word, supposedly unimportant variables. But by adding these other variables, AM is able to predict several behavioral properties of the indefinite article: (1) the one-way error tendency of adult speakers to replace *an* with *a* (but not *a* with *an*); (2) children’s errors favoring the extension of *a*, but not *an*, such as ‘a upper’, ‘a alligator’, ‘a end’, ‘a engine’, ‘a egg’, and ‘a other one’; (3) dialects for which *an* has been replaced by *a*, but not the other way around. In other words, the “unimportant” variables are crucial

for predicting the fuzziness of actual language usage (for some discussion of these properties, see Skousen 2003). Finally, another important property is that AM can predict the indefinite article even when the first segment is obscured (that is, when one cannot tell whether that segment is a consonant or a vowel). In such cases, the other variables are used to guess the syllabicity of the obscured segment, thus allowing for the prediction. In other words, AM allows for robustness of prediction. If we assume a symbolic rule system with only one rule (one based on the syllabicity of the first segment), then no prediction is possible when that segment is obscured. For additional discussion of the robustness of AM with respect to the indefinite article, see Skousen 1989: 58–9.

Specifying “unimportant” variables also allows for cases where the preferred analogy is not a nearest neighbor to the given context, but is found in a gang of homogeneous behavior at some distance from the given context. An important example of this occurs in predicting the past tense for the Finnish verb *sortaa* ‘to oppress’. Standard rule analyses of Finnish as well as nearest neighbor approaches to language prediction argue that the past tense for this verb should be *sorsi*, whereas in fact it is *sorti*. Yet when AM is applied to predicting the past tense in Finnish, it is able to predict the correct *sorti*, mainly because AM indirectly discovers that the *o* vowel is the “crucial” variable in predicting the past tense for this verb. In previous analyses (typically based on the historically determined “crucial” variables), the *o* vowel was ignored. But AM, by specifying variables (both “important” and “unimportant”) across the whole word, was able to make the correct prediction for this “exceptionally behaving” verb. For a complete discussion of how AM solves the problem of *sortaa*, see Skousen, Lonsdale, and Parkinson 2002: 27–36.

8.4.1 *Varying the granularity of prediction*

Computationally, there is a need to limit the number of variables. The current AM program can handle up to sixty variables, although the processing times can become quite long whenever there are more than forty variables. The problem here is that the actual computer program is sequential and does not simultaneously run an exponential number of cases (as the proposed quantum computer would). Even the parallel processing provided by standard supercomputers does not appear to be capable of eliminating the fundamental exponential explosion inherent in AM. Presumably there are also empirical limitations on the number of variables that are processed. In other words, there will be a degree and type of granularity that results from how many and which variables are selected. Ultimately, we have to select the variables, but we

want to judiciously select variables in a principled way that will, at the same time, allow for general applicability. In Skousen 1989: 51–4, I suggest that enough variables be selected so that each exemplar in the dataset is distinguishable or recognizable. It is this property that argues for specifying more than the first segment of the following word in predicting the indefinite article *a/an*. Or in the case of *sortaa*, we specify variables across the entire word (thus including the *o* vowel). Another suggestion is that proximity to the outcome should be accounted for. For instance, in trying to predict the ending for a word, if we want to provide variables for the antepenultimate syllable, we should also provide variables for the penultimate and ultimate syllables.

8.4.2 *Avoiding inappropriate variables*

There are undoubtedly some variables that are inappropriate, either conceptually or empirically. For instance, in predicting the negative prefix for adjectives in English, we could consider specifying the etymological source of adjectives since there is some correlation (although imperfect) between selecting the Latin negative prefixes *in-*, *il-*, *ir-*, and *im-* for words of Latin origin and the invariant Germanic negative prefix *un-* for words of Germanic origin. It turns out that such an etymological variable will have some influence in helping to predict the correct prefix (but not as much as one might suppose since historically these prefixes have been extended to words of different etymological background). From a conceptual point of view, in modern English, we cannot claim that speakers know the etymologies of the adjectives (although this may have been true for some educated speakers earlier in English when the influx of Latin vocabulary was in its beginning). For further discussion of this issue, see Chapman and Skousen 2005: 341–2.

As an example of an empirical restriction on variables, consider whether multisyllable words should be specified in terms of stress pattern or number of syllables. For instance, in predicting the past tense for Finnish verbs, Skousen 1989: 101–4 used a restricted dataset: two-syllable verbs ending in a nonhigh, unrounded vowel (*e*, *ä*, or *a*). The results were very accurate in predicting speakers' intuitions as well as historical and dialect development. But extending the dataset to the entire verb system was much more difficult until it was realized that the variables should be specified in terms of stress pattern rather than by number of syllables. This difference may seem surprising since stress is supposed to be fully predictable in Finnish (primary stress on the first syllable, secondary stress on alternating syllables according to syllable weight). Yet there is empirical evidence that Finnish speakers rely on stress rather than number of syllables. Consider the following two analyses of

the Finnish illative ending (meaning ‘into’), where the first analysis is based on counting the number of syllables, the second on the kind of stress placed on the last syllable:

number of syllables

one syllable, long vowel or diphthong	-hV _i n
two or more syllables	
long vowel	-seen
diphthong	-hV _i n
short vowel	-V _i n

stress

stressed, long vowel or diphthong	-hV _i n
unstressed	
long vowel	-seen
diphthong	-hV _i n
short vowel	-V _i n

(Here V_i means that the stem-final vowel is copied.) There is basically no difference between these two analyses since primary stress is virtually always on the first syllable. The crucial distinction between the two analyses is brought out when we consider how Finnish speakers predict the illative for two-syllable loan words where the original primary stress on the final syllable has been maintained. And the answer is that they follow the stress-based analysis:

Rousseau	rusó:	rusó:hon
Bordeaux	bordó:	bordó:hon
Calais	kalé:	kalé:hen

But if these words were nativized, with stress on the first syllable, then speakers would produce illative forms like /kále:se:n/. This means that in specifying the variables for Finnish words, we need to provide information regarding the stress pattern, not the number of syllables.

8.4.3 *Weighting of variables*

Now if we decide that we must specify the stress pattern, an important question arises: What is the strength of the stress in predicting the outcome? Is it the same as the individual phoneme? Consider, for instance, variables that might be specified for the syllable in Finnish (the nine variables listed here are much like the ones used in Skousen 1989):

- 1 syllable-initial consonant (include *o* as a possibility)
 - * a syllable-structure alternative:
 - (1a) is there an initial consonant or not?
 - (1b) if so, what is it?
- 2 the nuclear vowel: specify its phoneme
- 3 is there a second vowel or not?
- 4 if so, what is it?
- 5 is there a sonorant in the coda?
- 6 if so, what is it?
- 7 is there a obstruent in the coda?
- 8 if so, what is it?
- 9 what is the stress on the syllable? primary, secondary, none
 - * a scalar alternative (10, 00, 01):
 - (9a) is the stress primary?
 - (9b) is there no stress?

If we follow the two alternatives (each marked with an asterisk), we have eleven variables, of which four deal with syllable structure, five specify the sounds (here the phonemes), and two the stress. If we analyze the phonemes into distinctive features, the number of variables specifying sounds would at least triple and probably overwhelm the analysis. Perhaps even as it is, the two variables dealing with stress may not be enough. Even worse would be specifying a single stress variable for the intonational contour of the entire word.

This problem becomes more acute when one specifies variables from completely different types of linguistic classification, say phonetic and semantic. Suppose we are trying to predict an outcome, say a grammatical gender, that is affected by the phonetics of the word as well as whether the word refers, say, to animates or nonanimates. We set up say ten or so variables for the phonetics of the last syllable (as a minimum). But then the question is: Do we assign just one variable to tell us whether the word refers to an animate or nonanimate object? It is very doubtful that a single variable assigned to animacy will be strong enough to show the influence of that semantic class. Just doubling or tripling that semantic variable seems awfully arbitrary, although from a pragmatic point of view one could increase the strength of such a variable until one gets the right results! David Eddington did precisely that when he considered the relative strength of phonemic variables versus morphological variables in predicting Spanish stress assignment (Eddington 2002: 148):

Therefore, in addition to the phonemic information, morphological variables were included. For verbal forms, one variable indicated the person, and three identical variables indicated the tense form of the verb. Repeating a variable more than once is the only way to manipulate the weight of one variable or another prior to running the

AM program. In essence, what this implies is that the tense form of the verb is considered three times more important than any single onset, nucleus or coda. In the AM simulation, the only significant difference that weighting this variable made was in the number of errors that occurred on preterit verbs with final stress. Fifty errors occurred without the weighting, in comparison to 27 when it was included three times.

And it should be remembered that this approach will not work if the variable being considered has no effect on the outcome. Dirk Elzinga 2006: 766 reports that, in using AM to predict the comparative for English adjectives, he used a morphological variable that specified whether the adjective was morphologically simple or complex, and he discovered that doubling, tripling, and quadrupling that morphological variable had no effect on the predictability of the outcome.

Obviously, we need a principled method of constructing variables so that the empirically determined relative strength between classificatory types is naturally achieved.

8.5 Specifying the outcomes

8.5.1 Combining outcomes

In making predictions, one has to specify what the outcomes are. The issue is whether we should consider two or more outcomes as different or as variants of the same outcome. Sometimes this issue involves cases of abstractness. For instance, in the Latin negative prefix *in-*, used in English, there are several variants that show up: *il-* for words beginning with *l* (such as *illegal*), *ir-* for words beginning with *r* (such as *irregular*), and *im-* for words beginning with labials (such as *impossible*). When trying to predict the negative adjectival prefix in English, do we consider these four variants as a single morphological outcome (say, the abstract *IN-*) or as four different ones (*in-*, *il-*, *ir-*, or *im-*)? In general, our decision will affect our predictions of the negative adjectival prefix, and from those results we can perhaps discover which treatment (one or four outcomes) best accounts for speakers' actual predictions. For further discussion, see Chapman and Skousen 2005: 12.

Another example of this problem of outcome specification arises in the case of the Finnish illative ending *-hV_n* (discussed in Section 8.4.2). There we considered this ending as a single outcome, but theoretically one could consider it as a multitude of distinct outcomes, each different with respect to the copied vowel V_i :

voi 'butter'	voihin	-hin
syy 'reason'	syhyhyn	-hyn
kuu 'moon'	kuuhun	-hun
tie 'road'	tiehen	-hen
työ 'work'	työhön	-hön
suo 'swamp'	suohon	-hon

pää 'head'	päähän	-hän
maa 'land'	maahan	-han

Again, the issue is empirical; and the best predictions occur if we treat all of these forms as the same outcome, not as eight distinct outcomes (the latter leads to a substantial increase in the heterogeneity of the contextual space and subsequent loss in predictability). Such an analysis argues that speakers are therefore aware of the basic identity of all these variant forms.

8.5.2 *Separating or combining the outcomes*

Another issue deals with whether we have a single outcome or separate outcomes that apply in some order with respect to each other (or perhaps independently of each other). As an example of this, consider plural formation in German. The plural form can be viewed as two processes, adding an ending and mutating the stressed stem vowel (umlauting):

<i>singular</i>	<i>plural</i>	<i>ending</i>	<i>umlauting</i>
Berater 'advisor'	Berater	Ø	no
Vater 'father'	Väter	Ø	yes
Bauer 'farmer'	Bauern	n	no
Motor 'motor'	Motoren	en	no
Tag 'day'	Tage	e	no
Band 'volume'	Bände	e	yes
Band 'ribbon'	Bänder	er	yes
Band 'bond'	Bande	e	no
Band 'band'	Bands	s	no

One issue here is whether *-n* and *-en* should be considered syllabic variants of the same ending. Another issue involves the case when the stressed stem vowel is already a front vowel; in that case, we may ask whether one should consider umlauting as vacuously applying or not at all:

<i>singular</i>	<i>plural</i>	<i>ending</i>	<i>umlauting</i>
Rücken 'back'	Rücken	Ø	yes or no?
Bild 'picture'	Bilder	er	yes or no?
Bär 'bear'	Bären	en	yes or no?
Brief 'letter'	Briefe	e	yes or no?

Ultimately, the issue is how tightly linked are the endings with the umlauting. For some endings (such as *-er*), we expect umlauting (whenever it can apply). For other endings (such as *-en* or *-s*), we do not expect umlauting (whenever it can apply). And for some endings (such as *-e*) we can have umlauting or not, depending on the word (and again, whenever it can apply). These links between the ending and umlauting suggest that we should consider the cases of plural

formation as single outcomes. But ultimately, the issue is empirical. For instance, when the stressed stem vowel is not already a front vowel, do speakers (in the historical or dialectal development of the language or as children learning the language) remove the umlauting for the *-er* ending (which expects umlauting whenever it can apply)? If so, then we may wish to predict the ending and the umlauting separately from one another – or perhaps sequentially, with one being predicted first, then the other being predicted on the basis on the first prediction.

This example brings up the more paramount question of sequential versus simultaneous prediction in dealing with syntactic prediction and, we should add, virtually every other kind of linguistic prediction. Language processing involves sequencing through time, with one prediction following another and typically depending on previous decisions.

8.6 Repetition in the dataset

The final issue that I would like to bring up here is the question of how exemplars should be represented in the dataset. In Skousen 1989, I almost always listed the exemplars for morphological problems as types rather than as tokens. And in most instances, types have worked much better than tokens in predicting morphological behavior. When tokens are specified, the highly frequently occurring types typically overwhelm the analysis. In Skousen 1989: 54, I discuss the issue of types versus tokens and observe that “ultimately, the difference between type and token can be eliminated by specifying enough variables. By increasing the number of variables every token occurrence will also represent a single type”. But whether this proposal is feasible is questionable since there is undoubtedly some empirical limitation on the number of variables that can be handled.

The need to distinguish between types and tokens in phonetic and morphological problems has been emphasized in Bybee 2001: 96–136. Baayen and his colleagues (see de Jong, Schreuder, and Baayen 2000) have been arguing that a more accurate exemplar basis would be family types, where datasets would list all the morphologically related types, both inflectional and derivational, in datasets. Again, decisions of this sort regarding what to put in the dataset is an empirical issue.

8.7 Acknowledgments

I wish to thank members of the Analogical Modeling Research Group at Brigham Young University for their helpful criticisms and suggestions for improvement: Deryle Lonsdale, David Eddington, Dirk Elzinga, Dana Bourgerie, and Don Chapman. I also wish to thank Benjamin Skousen for his comments on the Chinese classifiers.

Modeling analogy as probabilistic grammar^{*}

Adam Albright

9.1 Introduction

Formal implemented models of analogy face two opposing challenges. On the one hand, they must be powerful and flexible enough to handle gradient and probabilistic data. This requires an ability to notice statistical regularities at many different levels of generality, and in many cases, to adjudicate between multiple conflicting patterns by assessing the relative strength of each, and to generalize them to novel items based on their relative strength. At the same time, when we examine evidence from language change, child errors, and psycholinguistic experiments, we find that only a small fraction of the logically possible analogical inferences are actually attested. Therefore, an adequate model of analogy must also be restrictive enough to explain why speakers generalize certain statistical properties of the data and not others. Moreover, in the ideal case, restrictions on possible analogies should follow from intrinsic properties of the architecture of the model, and not need to be stipulated post hoc.

Current computational models of analogical inference in language are still rather rudimentary, and we are certainly nowhere near possessing a model that captures not only the statistical abilities of speakers, but also their preferences and limitations.¹ Nonetheless, the past two decades have seen some key advances. Work in frameworks such as neural networks (Rumelhart and McClelland 1987; MacWhinney and Leinbach 1991; Daugherty and Seidenberg 1994, and much subsequent work) and Analogical Modeling of

^{*} Thanks to Jim Blevins, Bruce Hayes, Donca Steriade, participants of the Analogy in Grammar Workshop (Leipzig, September 22–3, 2006), and especially to two anonymous reviewers, for helpful comments and suggestions; all remaining errors are, of course, my own.

¹ Recent decades have seen a wealth of frameworks for modeling analogical inference and decision making more generally; see especially Gentner, Holyoak, and Kokinov (2001) and Chater, Tenenbaum, and Yuille (2006).

Language (AML; Skousen 1989) have focused primarily on the first challenge, tackling the gradient of the data. This work has had several positive influences on the study of analogy, particularly as a synchronic phenomenon. First, it has fostered a culture of developing computationally implemented models. These allow for objective tests of the extent to which a particular pattern can be extracted from the training data, given an explicitly formalized set of assumptions. In a few cases, such work has even led to implemented models of analogical change over time (e.g., Hare and Elman 1995). More generally, it has inspired a good deal of empirical work probing the detailed statistical knowledge that native speakers have about regularities and subregularities surrounding processes in their language. The overall picture that has emerged from such work is one of speakers as powerful statistical learners, able to encode a wide variety of gradient patterns.

In this chapter, I will take on the latter side of the problem, which has so far received far less attention in the literature: why do speakers generalize some regularities and not others? I discuss three general restrictions on analogical inference in morphophonology. The first is a restriction on how patterns are defined, which distinguishes between patterns that can be noticed and extended, and those that are evidently ignored. The second is a restriction on how patterns are evaluated, and concerns what it means for a pattern to be “well attested” or strong enough to generalize to novel items. The last is a restriction on which forms in a morphological paradigm are open to analogical change, and what determines the direction of influence. I argue that in all three cases, the observed restrictions correspond to limitations imposed by formulating processes as SPE-style rewrite rules ($A \rightarrow B / C_D$). This observation is not a trivial one, since this rule notation is a very particular hypothesis about how linguistic knowledge is structured, and how it makes reference to positions, variables, and so on. I demonstrate ways in which statistical models that lack this type of structure suffer in their ability to model empirical data, by overestimating the goodness of various possible but unattested types of analogical inference. Based on this observation, I argue that the best formal model of analogy is one that adds a probabilistic component to a grammar of context-sensitive statements.

The outline of the chapter is as follows: for each of the three proposed restrictions, I first present empirical data illustrating how it distinguishes attested from unattested analogies. Then, I compare two representative models, one with and one without the restriction imposed by rule-like structure. Finally, I discuss the broader implications of these observations for formal models of analogy.

9.2 What is a linguistically significant pattern?

9.2.1 *Structured vs unstructured inference*

To illustrate the role that a formalism can play in restricting possible analogies, it is instructive to start by considering the most traditional of all formalisms: four-part analogy. In four-part notation, analogies are expressed in the form in (1):

- (1) Four-part notation: $A:B :: X:Y$
 “Whatever the relationship is between A and B , it should also hold between X and Y ”

Discussions of four-part analogy frequently point out that the relation between words A and B is in many cases part of a much more general pattern, and that the examples A and B should be construed as representative members of a larger analogical set, consisting of more words ($A_1:B_1 :: A_2:B_2 :: A_3:B_3 :: \dots$) and perhaps also more paradigmatically related forms ($A_1:B_1:C_1 :: A_2:B_2:C_2 \dots$). The notation itself does not provide any way to indicate this fact, however, and thus has no formal means of excluding or disfavoring analogies supported by just one or a few pairs. Furthermore, the notation does not impose any restrictions on what properties particular $A_i:B_i$ pairs can have in common with one another. In fact the pattern itself—i.e., the relation between A and B , and the equation for Y —is left entirely implicit. This means that there are many possible ways to construct analogical sets, and few concrete ways to compare competing analogical inferences.

As an example, consider mid vowel alternations in Spanish present-tense indicative verb paradigms. In some verbs, when the mid vowels /e/ and /o/ are stressed, they irregularly diphthongize to [jé] and [wé], respectively. This occurs in the 1sg, 2sg, 3sg, and 3pl (as well as the entire present subjunctive). In other verbs, the alternation does not occur, and invariant mid vowels or diphthongs are found throughout the paradigm.

- (2) Spanish present tense diphthongization

a. Diphthongizing verbs

Verb stem	Infin.	3sg pres. indic.	Gloss
sent-	sent-ár	sjént-a	‘seat’
kont-	kont-ár	kwént-a	‘count’

b. Nonalternating verbs

Verb stem	Infin.	3sg pres. indic.	Gloss
rent-	rent-ár	rént-a	'rent'
mont-	mont-ár	mónt-a	'ride/mount'
orjent-	orjent-ár	orjént-a	'orient'
frekwent-	frekwent-ár	frekwént-a	'frequent'

Since diphthongization is lexically idiosyncratic, Spanish speakers must decide whether or not to apply it to novel or unknown words. For example, if a speaker was faced with a novel verb [lerrár], they might attempt to construct analogical sets that would support a diphthongized 3sg form [ljérra]. Using the four-part notation, there are numerous ways this could be done, including:

(3) Analogical set 1:

$$\left. \begin{array}{l} \textit{errar} : \textit{yerra} \\ \textit{enterrar} : \textit{entierra} \\ \textit{aserrar} : \textit{asierra} \\ \textit{aferrar} : \textit{afierra} \\ \textit{cerrar} : \textit{cierra} \\ \dots \end{array} \right\} :: \textit{lerrar} : \textit{lierra}$$

(4) Analogical set 2:

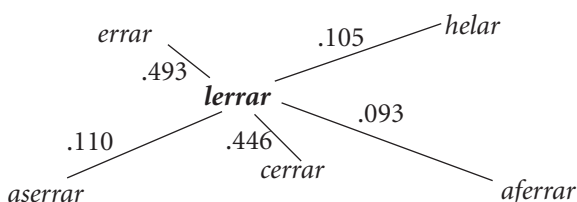
$$\left. \begin{array}{l} \textit{serrar} : \textit{sierra} \\ \textit{alentar} : \textit{alienta} \\ \textit{helar} : \textit{hiela} \\ \textit{querer} : \textit{quiere} \\ \dots \end{array} \right\} :: \textit{lerrar} : \textit{lierra}$$

The first set looks more convincing, since all of its members rhyme with [lerr-] and belong to the *-ar* inflectional class. Intuitively, this provides better support for the outcome [ljérra] than set 2 does; however, such a high degree of similarity is neither required nor rewarded by the formalism. In addition, nothing formally rewards a larger set (a point we will return to below). In sum, while the generality and flexibility of four-part notation have made it a convenient tool for describing analogical changes, as is often noted in the

literature, an explanatory theory of analogy depends on being able to impose restrictions on possible proportions (Morpurgo Davies 1978).

Let us start by addressing the first shortcoming of four-part notation, namely, its inability to capture the relative similarity of different analogical pairs to the target word. A common intuition about analogical sets is that they are not chosen randomly from the lexicon at large, but rather should represent the words that are expected to have the greatest influence because they are phonologically most similar to the target word—i.e., the closest analogs. For example, the existing Spanish verbs that are most similar to the novel verb [lerrar] are shown in (5) (similarity values are in arbitrary units, higher = more similar):

- (5) Existing Spanish verbs similar to [lerrar]



The restriction that we want the model to obey, then, is that generalization of a pattern to novel items must be supported by sufficiently many close analogs. One obvious way to do this is to adopt a similarity-based classification model, which decides on the treatment of novel items by considering its aggregate similarity to the set of known items. In such a model, the advantage of being similar to many existing words is anything but accidental; it is built in as a core principle of the architecture of the model.

There are many ways to be similar, however, and it is an empirical question what types of similarity matter most to humans in deciding how to treat novel words. For instance, the existing Spanish verbs *errar* ‘err’ and *cerrar* ‘close’ are similar to novel *lerrar* by ending in root-final [err]. The verb *helar* ‘freeze’ is also (at least somewhat) similar to *lerrar*, but this is due to the shared [l] (or perhaps the similarity of [l] and [r]), a similar syllabic structure, and so on. Hypothetical verbs like *lerdar*, *lenar*, and *lorrar* also share commonalities with *lerrar*, but each in its own unique way. Looking back at analogical set 1 in (3), there are intuitively two factors that make this group of analogs seem particularly compelling. First, all of these verbs share a set of common properties with each other and with the target word: they all end in [err] and all belong to the *-ar* inflectional class. In addition, those shared properties are

perceptually salient (involving rhymes of stressed syllables), and are local to the change in question (being either in the same syllable as the stressed mid vowel, or in the adjacent syllable). Albright and Hayes (2003) refer to this situation, in which the comparison set can be defined by their shared properties, as *STRUCTURED SIMILARITY*. If we compare analogical set 2 in (4), we see that *serrar*, *alentar*, *helar*, and *querer* share no such properties.² Albright and Hayes refer to this as *VARIEGATED SIMILARITY*.

Not all similarity-based models care about the exact source or nature of similarity. In principle, the similarity of the novel word to each existing word could be calculated independently. (An example will be given in the next section.) In order to give preference to structured similarity, a model must be able to align words with one another, determine what they have in common, and ignore what is unique to individual comparisons. This requires that the model have the capacity to encode the fact that a number of words all have the same type of element in the same location—that is, the model must be able to impose structure on the data, and encode its knowledge in terms of these structures (features, prosodic positions, etc.). This sounds like a simple requirement, but in fact it represents a fundamental divide between two classes of models: those that generalize using “raw” (unstructured) similarity to known words, and those that generalize by imposing structure on novel items and parsing them for elements in common with known words.

The goal of the rest of this section is to show that structured similarity is an important component in modeling how speakers generalize morphophonological patterns. The strategy will be as follows: first, in Sections 9.2.2–9.2.3, I will present two computationally implemented models, one lacking structured representations, and one that encodes its knowledge in structural terms. Then in Section 9.2.4, the performance of the two models will be compared against experimentally obtained data concerning the relative likelihood of different novel Spanish verbs to undergo diphthongization. To preview the results, it will emerge that the ability to make use of variegated (unstructured) similarity turns out to be not only unnecessary, but even harmful in modeling human intuitions.

9.2.2 *Analogy without structure: “pure” similarity-based classification*

To assess the contribution of structured similarity to the performance of a model, we first need a baseline model that does not require structured comparisons. One commonly used model of similarity-based classification that has been widely applied in many domains is the *GENERALIZED CONTEXT*

² Or, more precisely, they share only very general properties which do not distinguish them from other verbs in the language, such as having a liquid, a stressable mid vowel, and so on.

MODEL (GCM; Nosofsky 1986, 1990). For some applications in linguistics, see Johnson (1997), Nakisa, Plunkett, and Hahn (2000), and Albright and Hayes (2003). In this model, the treatment of a novel item is determined by calculating its similarity to classes of known items (exemplars). In deciding whether to assign a novel item i to a particular class c , the model compares item i to each existing member j of class c . The similarity of i to the entire class is a function of the summed similarities of each individual class member:

- (6) Similarity of novel item i to class c (with members j) = $\sum e^{(-d_{i,j}/s)}$, where
- $d_{i,j}$ = the psychological distance between i and j
 - s = sensitivity (a free parameter of the model)

The probability of actually treating i as a member of class c is simply proportional to its similarity to the individual members:

- (7) Probability of assigning item i to class c = $\frac{\text{Similarity of } i \text{ to } c}{\text{Total similarity of } i \text{ to all classes}}$

This model is based on the premise that analogical sets are more compelling when they contain more members, and when those members are more similar to the novel item. In this way, the model satisfies the restriction that analogical generalization must be sufficiently supported by known items. The model does not place any inherent restrictions on the nature of the similarity relations, however, specifying only that it reflect some generic notion of the *psychological distance* between two words. At its simplest and most neutral, this would simply be their *perceptual distance*, or some holistic measure of how similar the words sound. Intuitively, words sound similar to one another if their component segments are similar—that is, if the sounds of one word are well-matched to those of the other. In order to calculate this, we need perceptual similarity values for arbitrary pairs of sounds, and also a method of determining the optimal alignment of sounds, given their similarities.

One technique for estimating the similarity of pairs of segments is to consider how many natural classes they both belong to. Frisch, Pierrehumbert, and Broe (2004), following Broe (1993) and Frisch (1996), propose the following ratio:

- (8) Similarity of sounds s_1, s_2 = $\frac{\text{Number of shared natural classes}}{\text{Number of shared} + \text{unshared natural classes}}$

Given these similarity values, an optimal alignment of the sounds in two words is one in which they can be transformed into one another in as few steps as possible (Bailey and Hahn 2001; Hahn, Chater, and Richardson 2003). This can be calculated by finding the minimum string edit (Levenshtein)

distance (Kruskal 1983); see Bailey and Hahn (2001) and Albright and Hayes (2003) for details of how this is implemented based on segmental similarity. The result is a score for each pair of words, reflecting the degree of similarity between corresponding segments and the extent of mismatches (noncorresponding material). For example, the similarity of the novel verb *lerrar* to the existing Spanish verb *errar* is calculated to be 0.493 (in arbitrary units), while the similarity of *lerrar* to *reglar* is 0.268, and to *lograr* is 0.203.

We can use this model to calculate the likelihood of diphthongizing a novel Spanish verb, by simply comparing the aggregate similarity of that verb against the set of existing diphthongizing and nondiphthongizing verbs. For example, the summed similarity of the novel verb *lerrar* to diphthongizing verbs is 4.936 (again, in arbitrary units), with the top contributors including verbs like *errar* (0.493), *cerrar* (0.446), *aserrar* (0.110), *helar* (0.105), and *aferrar* (0.093). The summed similarity of *lerrar* to nondiphthongizing verbs is 15.551, with top contributors including *reglar* (0.268), *orlar* (0.240), *ahorrar* (0.213), *ferrar* (0.211), and *lograr* (0.203). We see that the higher score for the nondiphthongizing comes not from greater similarity of any individual member—in fact, *errar* and *cerrar* in the diphthongizing class are much more similar than any nondiphthongizing verb. Rather, this advantage is due to the fact that there are many more nondiphthongizing verbs, so small amounts of moderate similarity sum up to outweigh a small number of very similar verbs. Using the equation in (7), the overall probability of applying diphthongization to *lerrar* is predicted to be $4.936 / (4.936 + 15.551)$, or 24.09%.

There are a couple points to note about the workings of this model. First, the model has the ability to make use of variegated similarity, since similarity is based on the optimal alignments of individual pairs of items. However, the examples in the preceding paragraph show that not all inferences make equal use of it; in fact, the closest analogs supporting diphthongization almost all contain *-errar*. This turns out to be quite typical, and analogical sets are frequently dominated by words that all happen to share the same feature(s) in common with the target word—i.e., a structured similarity. This aspect of the model will be important to keep in mind when evaluating the performance of the GCM, since we are interested not only in how well the model does, but also in the question of whether it benefits from its ability to use variegated similarity.

9.2.3 Analogy with structure: Probabilistic context-sensitive rules

As noted above, an ability to refer to particular properties of words (having a certain type of sound in a certain location, having particular prosodic properties, etc.) is crucial in requiring that analogical sets share structural

similarities. In fact, many modeling frameworks use structural properties to decide how to treat novel items. Feature-based classification models (Tversky 1977), such as TiMBL (Daelemans, Zavrel, Van der Sloot, and Van den Bosch 2000) and AML (Skousen 1989) directly incorporate the idea that in order for a group of items to be similar, they must share certain properties (feature values). Linguistic rules impose an even more specific structure. For example, context-sensitive readjustment rules ($e \rightarrow je / X _ \text{rro}]_{1\text{sg}}$) specify a change location, immediately adjacent left and right contexts, precedence relations, and so on. Although rule application is often thought of as fundamentally different from (and incompatible with) analogical inference, in fact, it is possible to think of rules as a very specific theory of how analogical sets are constructed—namely, by picking out groups of words that can be captured using the rule notation format.

The MINIMAL GENERALIZATION LEARNER (MGL; Albright and Hayes 2002) is a computationally implemented model that finds rules covering sets of words that behave consistently (belong to the same inflectional class, share the same morphophonemic change, etc.). It employs a bottom-up inductive procedure to compare pairs of words in the input data, find what they have in common, and encode these commonalities using a grammar of stochastic rules. For details of the model, the reader is referred to Albright and Hayes (2002) and Albright and Hayes (2003); in this section I provide a brief overview.

The model takes as its input pairs of forms that stand in a particular morphological relation, such as present/past, or infinitive/1sg, as in (9). In the present case, the relation between diphthongized and nondiphthongized stem variants is conditioned by stress placement, rather than any particular morphological category. Therefore, in the simulations reported here, input data are represented as pairs of stressed and stressless stem allomorphs, abstracting away from the suffixal material of the particular inflected forms that require one or the other, but retaining an indication of inflection class information (*-ar*, *-er*, *-ir*).

(9) Input to the minimal generalization learner: Some sample *-ar* verbs

Stressless	Stressed	Gloss	Orthography (infinitive)
jeg	jég	'arrive'	(<i>llegar</i>)
dex	déx	'leave'	(<i>dejar</i>)
jeb	jéb	'bring'	(<i>llevar</i>)
ked	kéd	'stay'	(<i>quedar</i>)
enkontr	enkwéñtr	'find'	(<i>encontrar</i>)

(continued)

(9) (*cont.*)

Stressless	Stressed	Gloss	Orthography (infinitive)
pens	pjéns	‘think’	(<i>pensar</i>)
kont	kwént	‘tell, count’	(<i>contar</i>)
entr	éntr	‘enter’	(<i>entrar</i>)
tom	tóm	‘take’	(<i>tomar</i>)
kre	kré	‘create’	(<i>crear</i>)
empes	empjés	‘start’	(<i>empezar</i>)
esper	espér	‘wait, hope’	(<i>esperar</i>)
rekord	rekwérd	‘remember’	(<i>recordar</i>)
tembl	tjémbl	‘tremble’	(<i>temblar</i>)

The first step in learning is to analyze individual (stressless, stressed) pairs, by factoring them into changing and unchanging portions. This allows each pair to be expressed as a rule, encoding both the change ($A \rightarrow B$) and the non-changing portion ($C _ D$). For example, the pair (tembl, tjémbl) has a vowel change surrounded by unchanging consonants: $e \rightarrow j\acute{e} / t _ mbl$ (“stressless [e] corresponds to stressed [je] when preceded by [t] and followed by [mbl]”). The pair (jeg, jég) on the other hand differs only in stress: $e \rightarrow \acute{e} / j _ g$.

Once the input pairs have been recast as word-specific rules, they are compared to find what they have in common, according to the rule scheme in (10):

(10) Comparing *tembl-/tiembl-* ‘tremble’, *desmembr-/desmiembr-* ‘dismember’:

Residue	Shared feats	Shared segs	Change loc.	Shared segs	Shared feats
des	t m		—	mb mb	l r
X	[-syllabic -continuant]		—	mb	[-syllabic +sonorant +continuant +voice +coronal +anterior]

The comparison in (10) yields a very specific rule that retains all of the properties shared by *tembl-* and *desmembr-*, subject to the restriction that they can be encoded in the structural components of the rule. Shared material

is expressed in terms of phonological features, while unshared material is expressed as variables. By convention, unmatched material on the left side is collapsed into a variable called 'X', and material on the right into a variable 'Y'. When such comparisons are carried out iteratively across the entire dataset, however, much broader rules can emerge through comparison of diverse forms, while further comparison of similar forms will yield additional narrow rules. A small sample of the many possible rules that could be learned from a set of Spanish verbs is given in (11).

(11) Representative rules for Spanish verbs³

- i. $o \rightarrow wé / [+consonantal] _ rs$
- ii. $o \rightarrow wé / \begin{bmatrix} -continuant \\ -voice \end{bmatrix} r _ \begin{bmatrix} -continuant \\ -syllabic \end{bmatrix}$
- iii. $o \rightarrow wé / \begin{bmatrix} -syllabic \\ +consonantal \end{bmatrix} _ \begin{bmatrix} -syllabic \end{bmatrix}$
- iv. $o \rightarrow ó / \begin{bmatrix} -syllabic \\ -sonorant \\ +consonantal \end{bmatrix} _ \begin{bmatrix} -syllabic \\ +consonantal \\ -continuant \end{bmatrix}$
- v. $o \rightarrow ó / \begin{bmatrix} -syllabic \\ +voice \end{bmatrix} _ \begin{bmatrix} -syllabic \end{bmatrix}$
- vi. $o \rightarrow ó / _ \begin{bmatrix} -syllabic \end{bmatrix}$

These rewrite rules incorporate many types of structure that limit possible comparisons. Rules specify linear relations such as precedence and adjacency. This notation rules out many logically possible sets of words, such as those that all have a certain sound, but its location is variably either to the right or the left of the change. This particular procedure also compares words by starting immediately adjacent to the change and working outwards, meaning that the descriptions of the left and right-side contexts are limited to the local contexts.⁴ Rule notation also embodies a form of strict feature matching: rules apply if their structural description is met, and not otherwise. Finally, although SPE-style rewrite rules are written in a way that could theoretically make use of the full power of context-sensitive grammars, the rules employed by this model obey commonly observed conventions for phonological rewrite rules by referring to a fixed number of positions and applying noncyclically,

³ Since the implemented model uses linear (flat) phonological representations, stress is encoded here as a feature of the stressed vowel, rather than as a property of the syllabic context.

⁴ Ultimately, this is too strong an assumption, since contexts are sometimes nonlocal. For an attempt to extend this system to find nonlocal contexts, and discussion of some of the issues involved, see Albright and Hayes (2006).

and thus are restricted to expressing regular relations which can be captured with a finite state transducer (Johnson 1972; Kaplan and Kay 1994; Gildea and Jurafsky 1996). The system thus embodies a very strong form of structured similarity: all that matters is that words are the same in the relevant respect, and there are no penalties or rewards for additional similarities or differences.

Once all of the possible rules have been discovered, it remains to decide which dimensions of similarity the speaker should actually pay attention to. In order to do this, the rules are evaluated according to their accuracy in the training data. The RELIABILITY of a rule is defined as the number of cases that it successfully covers (its HITS), divided by the number of cases that meet its structural description (its SCOPE). Raw reliability scores are then adjusted slightly downward using lower confidence limit statistics, to yield a score called CONFIDENCE. This has the effect of penalizing rules that are based on just a small amount of data (a small scope). The confidence scores for the rules in (11) are shown in (12):

- (12) Representative rules for Spanish, evaluated (hits/scope \Rightarrow confidence)
- i. $o \rightarrow w\acute{e} / [+cons] _ rs$ 4/4 \Rightarrow .786
 - ii. $o \rightarrow w\acute{e} / \begin{bmatrix} -contin \\ -voice \end{bmatrix} r _ \begin{bmatrix} -contin \\ -syll \end{bmatrix}$ 6/8 \Rightarrow .610
 - iii. $o \rightarrow w\acute{e} / \begin{bmatrix} -syll \\ +cons \end{bmatrix} _ \begin{bmatrix} -syll \end{bmatrix}$ 68/545 \Rightarrow .116
 - iv. $o \rightarrow \acute{o} / \begin{bmatrix} -syll \\ -sonor \\ +cons \end{bmatrix} _ \begin{bmatrix} -syll \\ +cons \\ -contin \end{bmatrix}$ 101/106 \Rightarrow .934
 - v. $o \rightarrow \acute{o} / \begin{bmatrix} -syll \\ +voice \end{bmatrix} _ \begin{bmatrix} -syll \end{bmatrix}$ 19/22 \Rightarrow .795
 - vi. $o \rightarrow \acute{o} / _ \begin{bmatrix} -syll \end{bmatrix}$ 588/668 \Rightarrow .871

Finally, the grammar of rules can be used to generalize patterns to novel items. The probability of generalizing a process is defined as in (13). Since this calculation is intended to mimic the probability with which a particular pattern will be employed to produce a target output, it is referred to as the PRODUCTION PROBABILITY of that pattern:

- (13) Production probability

$$= \frac{\text{Confidence of the best rule applying the pattern to the input}}{\text{Summed confidence of best rules applicable to the input, each pattern}}$$

For example, in calculating the likelihood to diphthongize the novel verb *lerrar*, the best (= most confident) applicable diphthongization and non-diphthongization rules are:

(14) Likelihood to diphthongize *lerrar*

- Best applicable diphthongization rule:

$$e \rightarrow \acute{e} / \left[\begin{array}{l} +\text{consonantal} \\ +\text{coronal} \end{array} \right] \text{ — } \left[\begin{array}{l} +\text{consonantal} \\ +\text{voice} \end{array} \right]$$

Reliability = 10/29; Confidence = .290

- Best applicable nondiphthongization rule:

$$e \rightarrow \acute{e} / \left[\begin{array}{l} -\text{syllabic} \\ +\text{voice} \end{array} \right] \text{ — } [+ \text{sonorant}]$$

Reliability = 86/86; Confidence = .989

- Production probability (*lirro*) = $\frac{.290}{.290 + .989} = 23\%$

For both the Minimal Generalization Learner and the Generalized Context Model, support for generalizations comes from large numbers of words that are similar to the target word and behave consistently. In the MGL, however, similarity is defined (in boolean fashion) as presence of certain structural features. This prevents the model from using variegated similarity, since such diverse sets of relations cannot be captured in the rule notation. We can contrast this with the GCM, in which the supporting words need not be similar to one another in any particular way. This leads to the possibility that analogical inference may be based on variegated support. In the next section, we attempt to test whether this additional ability is helpful or harmful to the GCM.

Finally, it is worth noting that proportional analogy is most often used in a way that conforms to the structural restrictions imposed by the rule-based model, since the antecedent in four-part notation requires that there is a well-defined relation, and ideally also a group of words that all share the same relation. Although individual analysts may disagree about what constitutes a valid relation (see Morpurgo Davies 1978 for a review of some prominent points of view), in practice, relations are most naturally thought of as a single rewrite relation, much as in SPE-style rules. This is not to say that the formalisms are equivalent, however, since proportional analogy is certainly flexible enough to encompass relations that cannot be expressed in rule-based terms. For example, nothing formally precludes setting up proportions showing relations that involve multiple changes (prefixation of [s] and nasalization of final consonant: *tick:sting :: crab:scram :: cat::scan?*), or changes that depend on the presence of an element somewhere in the word regardless of linear order (change of [ɪ] → [ʌ] adjacent to a [p]: *pinch::punch :: sip::sup :: pig:pug?*). A hypothesis of the rule-based model is that in order for a relation to be linguistically active—i.e., extended systematically

to new forms—it must involve a change defined in terms of phonological features, applied to a set of words that share a common structure (again, defined over linearly arranged combinations of natural phonological classes).

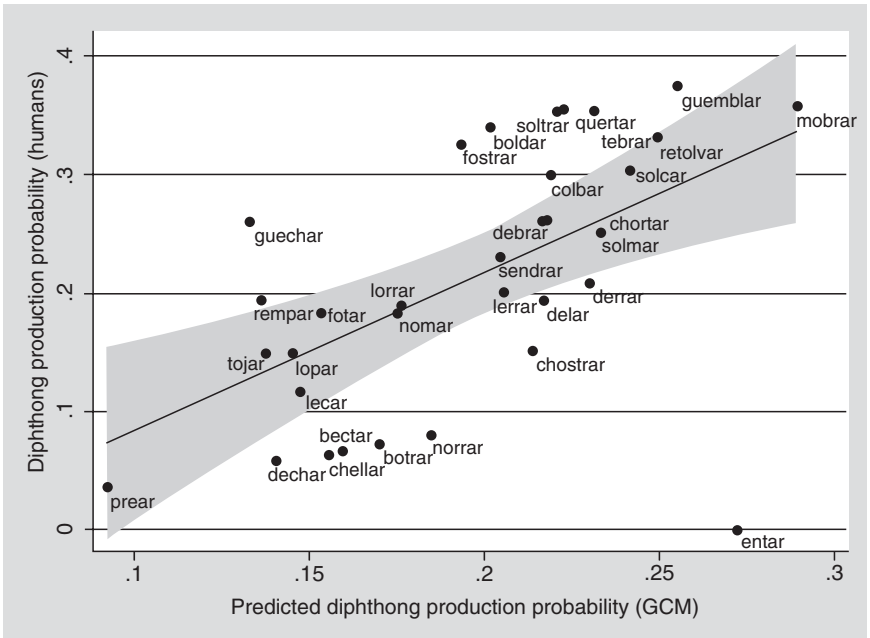
9.2.4 *An empirical test: Modeling diphthongization in novel words*

In order to test whether humans are restricted to inferences based on structured similarity, we can compare the performance of the two models against experimentally obtained data in which Spanish speakers were likewise tested on how they would produce stressed forms of novel verbs. Albright, Andrade, and Hayes (2001) asked 96 native speakers to inflect novel verbs containing mid vowels, to measure the relative likelihood of diphthongized responses in different contexts. Participants were given novel verbs in an unstressed form (e.g., [lerrámos] ‘we *lerr*’) and were asked to produce a stressed form (e.g., [lerro]/[ljérrro] ‘I *lerr*’). For each verb, the production probability of diphthongization was calculated by dividing the number of diphthongized responses by the total number of diphthongized + undiphthongized responses. For example, for the verb *lerrar*, 19 participants volunteered [ljérrro] and 76 volunteered [lérrro],⁵ yielding a 20% production probability of diphthongization. (For additional details of the experimental design and results, see Albright, Andrade, and Hayes 2001.)

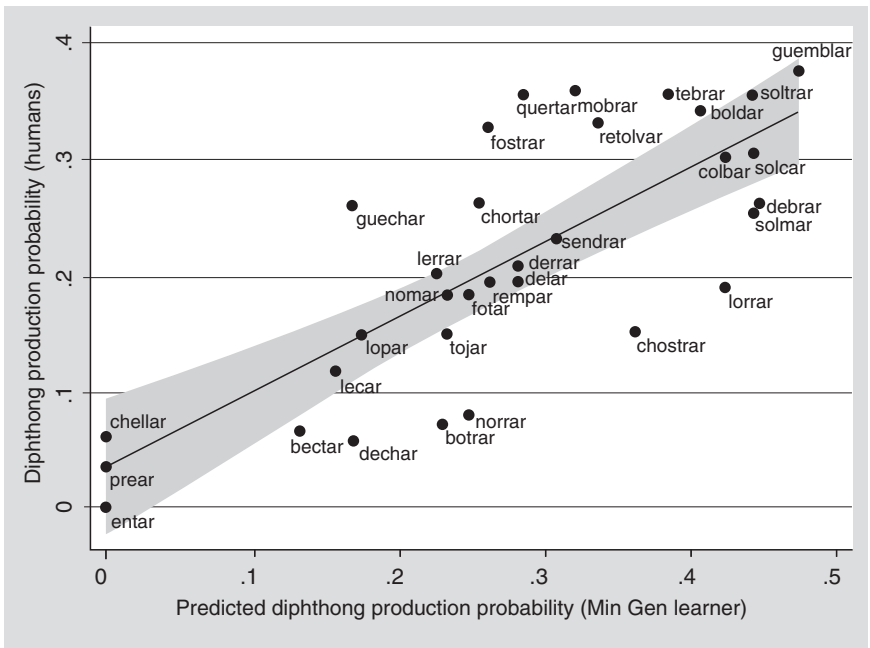
In order to test the models, predictions were obtained by training each model on a lexicon of Spanish. Two different datasets were tested: one that included all of the verbs in the LEXESP corpus containing stressable mid vowels (1,881 verbs total), and another that included just the subset of verbs that fall in the *-ar* inflectional class (1,669 of the total set). The choice of dataset turns out to matter slightly for the results, with the GCM performing slightly better on the full set and the MGL performing slightly better on the smaller set. The differences were relatively small, however, and I simply report here the better result for each model (i.e., treating the choice of dataset as a parameter that can vary independently across models).

Figure 9.1 shows the overall ability of the two models to predict the probability of diphthongization on a verb-by-verb basis. We see that both models do reasonably well, though the MGL does somewhat better ($r = .77$) than the GCM ($r = .56$). Most of this difference comes from the exceptionally poor performance of the GCM on a single outlier, however (*entar*); if this one item is excluded, the performance of the GCM is approximately as good as the MGL (r increases to $.74$).

⁵ One additional subject volunteered an unexpected and idiosyncratic change for this verb; this response was excluded.



a. Generalized Context Model ($r = .56$)



b. Minimal Generalization Learner ($r = .77$)

FIGURE 9.1 Predicted vs observed production probability of diphthongization

So what do we conclude from this result? Clearly, neither model can be rejected outright based on raw performance. In fact, the predictions of the two models are also significantly correlated with one another ($r = .53$). This means that the models are not merely making equivalently good predictions—in fact, to a large extent they are making the very same predictions.⁶ When the outputs of the two models are inspected, the reason is not hard to find: in very many cases, the two models pick out overlapping analogical sets. For example for the novel word *solmar*, the MGL found that the most confident applicable diphthongization rule was $o \rightarrow wé / s_ \ell Y]_{-ar \text{ class}}$ (including such words as *solar* ‘pave’, *soltar* ‘release’, and *soldar* ‘solder’). These same words figure prominently in the analogical set that the GCM employs; the five top contributors are *solar* (similarity .493), *soldar* (.417), *soltar* (.338), *cerrar* (.214), and *dormir* (.164). Similarly for the verb *lorrar*, the MGL used a rule $o \rightarrow wé / \left[\begin{array}{l} +\text{coronal} \\ +\text{continuant} \end{array} \right] _ \left[\begin{array}{l} +\text{coronal} \\ +\text{voice} \end{array} \right] Y]_{-ar \text{ class}}$

supported by positive examples like *solar*, *sonar*, *soldar*, *rodar*, and *soltar*. Here too, the rule includes three of the GCM’s five closest analogs: *errar* (.278), *cerrar* (.252), *solar* (.095), *rodar* (.094), and *soldar* (.085). The upshot is that although the GCM has access to variegated similarity—seen, for example, in the presence of analogs like *cerrar*—there is no guarantee that it is actually using it to a significant extent in any particular case. Thus, overall comparisons like the one in Figure 1 are unlikely to be illuminating about what mechanism speakers actually use to make analogical inferences.⁷

The examples in the preceding paragraph show that although in practice the role of variegated similarity is less than what is theoretically possible, the GCM does use it at least to a certain extent. What we need, then, is a way to focus specifically on the contribution of the variegated analogs, which the MGL cannot include as support for inferences. This requires a means of separating analogs that share a structured relation from those that do not. For a set like {*solar*, *soldar*, *soltar*, *cerrar*, *dormir*}, the intuitive division is between the first three, which share #*sol* (and the *-ar* inflectional class), as opposed to *cerrar* and *dormir*, which look like odd men out. Strictly speaking, however, it is not the case that these verbs completely lack structural properties with the remaining forms. In fact, all five verbs share the set of properties in (15):

⁶ This was confirmed by a stepwise multiple regression analysis, in which the MGL predictions were entered first with a high degree of significance ($p < .0001$), and the GCM predictions were unable to make any additional significant contribution.

⁷ The reason that the GCM tends to stick to such structurally interpretable analogical sets appears to be the fact that diphthongizing verbs in Spanish themselves happen to fall into such clusters. The

(15) Structural commonality: *solar*, *soldar*, *soltar*, *cerrar*, and *dormir*

$$\#[-\text{sonorant}] \begin{bmatrix} +\text{syllabic} \\ -\text{high} \\ -\text{low} \end{bmatrix} \begin{bmatrix} +\text{consonantal} \\ +\text{sonorant} \\ -\text{nasal} \end{bmatrix} \text{Y}$$

The description in (15) expresses a structured similarity, but expanding the context to include *cerrar* and *dormir* comes at a price. The description is now so general that it includes not just these five verbs, but also many others—including, importantly, some that do not diphthongize. In other words, although the description in (15) unifies all of the members of the GCM's analogical set, it does not accurately or uniquely describe what sets them apart from the rest of the verbs in the language. A rule-based model like the MGL could state a rule that applies diphthongization in this context, but it would not be a useful rule since it has too many exceptions.

This suggests a refinement to how we isolate sets of structured analogs: they must not only have in common a set of shared properties, but those properties must also be reliably associated with class membership. For example, it is not enough to be able to state what *cerrar* has in common with *soldar* and *soltar*; the properties that they share must also distinguish these verbs from nondiphthongizing verbs. In order to separate structured from unstructured analogs, then, we need a hypothesis about what those distinguishing properties are. Not coincidentally, this is precisely what the MGL model is designed to identify. For example, as noted above, the MGL determines that the properties of *solmar* that are most reliably associated with diphthongization are the preceding /s/ and the following /l/, making *solar*, *soldar*, *soltar* the analogs that share the set of most relevant structural properties. It should be possible, therefore, to use the structural descriptions that the MGL selects to help identify when the GCM is making use of unstructured, or variegated similarity.

explanation for this may be partly phonological, since phonotactic restrictions on stem-final consonant combinations would restrict the set of possibilities in this position, and make it easier for commonalities to emerge. There may also be a historical component: suppose the structured model of analogy is the correct one, and structure-guided inferences have been shaping Spanish over the centuries. In this case, we would expect verbs to retain diphthongization most readily if they fall into structurally definable gangs, creating structure in the lexicon of Spanish. If this were true, then the GCM could do good job of capturing the modern language, but would be unable to explain how the language came to be this way. If, on the other hand, the GCM model were correct, we would expect diphthongizing verbs to be retained on the strength of variegated similarity, and the set of existing diphthongizing verbs could consist of variegated analogical sets which the structured model would be unable to locate. A full diachronic analysis of verb-by-verb changes in diphthongization is left as a matter for future research.

In order to quantify the contribution of nonstructured analogs in the predictions of the GCM, I first ran the MGL, finding for each nonce form the set of properties that were found to be most reliably associated with diphthongization (i.e., the structural description of the best applicable rule that could derive a diphthongized output). I then ran the GCM, collecting the set of diphthongizing analogs. For each nonce verb, the analogical set was then separated into two groups: the structured analogs, which contained the best context identified by the MGL, and the variegated analogs, which fell outside this context. Examples for the novel verbs *solmar* and *lorrar* are given in (16).

(16) Separating structured vs variegated analogs

a. *solmar*: best context = [sol...]_{-ar class})

Structured analogs	Unstructured analogs
<i>solar</i> .493	<i>serrar</i> .214
<i>soldar</i> .417	<i>dormir</i> .164
<i>soltar</i> .338	<i>sonar</i> .157
	<i>serner</i> .139
	<i>socar</i> .126
	(and 235 others)

b. *lorrar*: best context = [$\begin{bmatrix} +\text{coronal} \\ -\text{continuant} \end{bmatrix}$ o $\begin{bmatrix} +\text{coronal} \\ +\text{voice} \end{bmatrix}$]_{-ar class}

Structured analogs	Unstructured analogs
<i>solar</i> .095	<i>errar</i> .278
<i>rodar</i> .094	<i>serrar</i> .252
	<i>soldar</i> .085
	<i>forsar</i> .084
	(and 236 others)

The contribution of variegated analogy was then defined as the summed similarity of the unstructured analogs divided by the summed similarity of all analogs (structured and unstructured). This ratio is taken as a measure of the extent to which the GCM is relying on variegated similarity for any particular nonce word.

We are now in a position to evaluate the usefulness of variegated similarity. If speakers make analogical inferences in a way that is blind to structure, then the MGL model should suffer in cases where variegated similarity is needed, since it is unable to make use of a crucial source of support. Conversely, if structure is critical to how speakers generalize, then the GCM should do

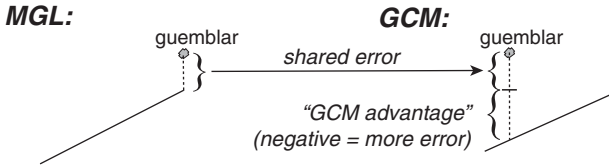


FIGURE 9.2 Calculation of “GCM advantage” score based on residuals

worse the more it relies more on variegated similarity. The word-by-word performance of each model was tested by fitting the predictions of each model against the experimentally obtained human responses using a linear regression. For each word, it was then determined how far off the model was, by subtracting the observed from the predicted values (i.e., calculating the residuals). The performance of the two models was collapsed into a single “GCM advantage” score by subtracting the GCM error from the MGL error for each word; this score is positive for a particular word if the MGL’s prediction is less accurate than the GCM’s, and negative if the GCM is farther off. This comparison is illustrated in Figure 9.2. Finally, the GCM advantage scores were correlated against the contribution of variegated analogy, as defined in the preceding paragraph. If variegated analogs are important to speakers, we expect a positive correlation, since in cases where variegated similarity plays a larger role, the MGL should suffer more (positive GCM advantage). If speakers do not use variegated similarity, we expect a negative correlation, since the GCM’s reliance on variegated analogs would encourage generalizations that humans do not make. In fact, when the correlation is calculated as described above, the result is weakly negative ($r = -.195$). Thus, we fail to find any support for the idea that variegated similarity is needed—and in fact, there is an indication that it may even be harmful.

The same result can also be seen another way, by calculating for each novel item the degree to which the GCM overestimated the goodness of each output. This amount will be positive if the GCM assigned too high a score (overpredicting the goodness of the output), and zero if the GCM is right on or under. The rationale for restricting the analysis to *overpredictions* is the following: suppose that speakers do not notice variegated similarity, and that the GCM is incorrect to use it. If this is true, then access to variegated analogs should let the GCM (incorrectly) gather extra support for some outputs, leading to overestimation of their goodness. Therefore, the negative effects of variegated similarity should be seen most clearly in the GCM’s overprediction errors. To test this, the GCM’s overestimation scores were correlated against the relative contribution of variegated similarity, as defined above. The result

here was a positive correlation between variegated similarity and overestimation ($r = .33$). This shows that the extra sources of support that the GCM has access to are not helpful in modeling speakers more accurately—in fact, they are deleterious, causing the model to overestimate the probability of diphthongization. Albright and Hayes (2003) also make a similar point about the GCM, using data from English past-tense formation.

There is finally one last way in which structure can be seen to matter. If we examine the GCM predictions in Figure 9.1a, we see that the most blatant gaffe by far that the GCM makes is in overpredicting the probability of diphthongization in *entar*. This prediction is based on the support of diphthongizing analogs like *sentar* ‘seat’, *mentar* ‘mention’, *tentar* ‘touch’, *dentar* ‘teethe’, *ventar* ‘sniff’, and so on. All of these analogs have a preceding consonant, and in fact diphthongization of initial vowels is overall quite rare in Spanish (particularly in the *-ar* class). The MGL is able to encode this fact by requiring that a consonant is a crucial part of the context when formulating rules. The GCM, on the other hand, has no way to encode this beyond the standard penalty for inserting or deleting a single segment in the process of calculating the optimal string alignment; therefore, it cannot categorically block analogy to similar consonant-initial words. This is yet another indication that speakers encode knowledge of patterns in terms of properties of elements that appear in particular positions—that is, in terms of linguistic structure.

9.2.5 Local summary

In this section, I have discussed a major restriction on what type of pattern can be generalized through analogy: it must be supported by sets of words that share a particular combination of properties in common, both with each other and with the target word. This may seem like an obvious or trivial restriction, and in fact many models simply assume it without argument. However, it is certainly not a logically necessary part of how analogy is formalized. Many exemplar-based models, such as the GCM, do not obey this restriction. This allows them to capture a wider range of patterns, and thereby makes them less constrained models. I have shown that the extra power afforded by unstructured comparisons does not help—and indeed, it seems to hurt by inflating the predicted goodness of certain generalizations. This confirms similar results shown previously for English by Albright and Hayes (2003).

Importantly, the restriction to structured comparisons is exactly what we would expect if speakers encode patterns using something like probabilistic

context-sensitive rules, of the sort employed by the MGL. Of course, this is not the only model that imposes structure on its representations; similar restrictions are also found in feature-based models, such as TiMBL (Daelemans, Zavrel, Van der Sloot, and Van den Bosch 2000) and AML (Skousen 1989).

9.3 Type vs token frequency

Another possible restriction on analogical models concerns the way in which the support for competing patterns is evaluated. In principle, a pattern could be strengthened in at least two different ways: by occurring in a large number of different words (high type frequency), or by occurring in a smaller number of words that are used very commonly (high token frequency). In fact, it appears that the propensity to generalize morphophonological patterns to new forms depends primarily on type frequency, and not on token frequency. This restriction has been noted numerous times in the literature; see Baayen and Lieber (1991) for English derivational suffixes, Bybee (1995) for French conjugation classes, German past participles, and others, Albright (2002*b*) for Italian conjugation classes, Albright and Hayes (2003: 133) for English past tenses, Ernestus and Baayen (2003: 29) for stem-final voicing in Dutch, Hay, Pierrehumbert, and Beckman (2004) for medial consonant clusters in English, and additional references in Bybee (1995). In this section, I provide further evidence for this conclusion, and suggest that it favors a model in which patterns are abstracted from individual words and encoded in some form that is separate from the lexicon (such as a grammar).

The formal definition of similarity in the Generalized Context Model ((6) above) is compatible with counting based either on type or token frequency, since “members of a class” could be taken to mean either types or individual tokens. In practice, however, the most natural interpretations of the model would lead us to expect a role for token frequency. If we assume, as is often done, that the GCM operates over exemplar representations (Johnson 1997; Pierrehumbert 2001), then every single token should contribute a measure of support to the strength of the pattern. Furthermore, even if we assume that the GCM operates over a more schematic lexicon that abstracts away from individual exemplars, there is ample evidence from online recognition and processing tasks that words with higher token frequency are accessed more readily than low-frequency words. Therefore, even if the GCM counts over a lexicon distinct word types, it seems likely that token frequency effects would emerge simply because of the way the lexicon is accessed. Stated more

generally, the premise of the GCM is that generalization is carried out by consulting the lexicon directly, and token frequency effects are characteristic—perhaps even diagnostic—of lexical access. It is important to note that the GCM is also very sensitive to type frequency, since each type contributes at least one token to the summed support for a particular class.

In principle, the Minimal Generalization Learner could also evaluate rules using types or tokens, but the rules it discovers are most naturally interpreted in terms of types. The comparisons that it carries out to abstract away from individual lexical items ((10) above) require just a single instance of each word, and nothing more can be learned from further tokens of previously seen data. In a system in which additional tokens are gratuitous, it would perhaps be a surprising design feature if token frequency played a crucial role in how rules are evaluated. In fact, calculating the confidence of rules according to their token frequency would require extra work in this model, since repeated tokens of the same lexical item could otherwise be disregarded as uninformative. This also relates to the more general hypothesis that grammars are intrinsically about kinds of words, rather than about particular instances of their use. Therefore, even if it is not strictly speaking required by the formalism, a rule-based account of analogy is most naturally limited to the influence of type frequency.

Spanish diphthongization provides a direct test of the relative importance of type vs token frequency, since although diphthongization is a minority pattern in the Spanish lexicon, affecting only a relatively small number of mid-vowel verbs (lowish type frequency), the verbs that undergo it tend to be among the most frequent verbs in the language (high token frequency). There is abundant *prima facie* evidence that the high token frequency of diphthongization does not make it a strong pattern: synchronically it is relatively unproductive in experimental settings (Bybee and Pardo 1981; Albright, Andrade, and Hayes 2001), and diachronically verbs tend to lose diphthongization alternations (Penny 2002; Morris 2005). Furthermore, overregularization errors among children acquiring Spanish consistently result in omitting diphthongization (Clahsen, Avelado, and Roca 2002), even though diphthongizing tokens constitute a large portion—perhaps even the majority—of children's experience.

In order to test the influence of token frequency more systematically, I ran both the GCM and the MGL with and without taking token frequency into account. Specifically, a weighting term was introduced in the GCM, so the contribution of each analog was defined not only in proportion to its similarity, but also in proportion to its (log) token frequency. A weighting term was also introduced into the MGL, such that the contribution of each word to the hits and/or scope of a rule was weighted according to its log token

frequency. The result was that both models did slightly worse when token frequency was taken into account, as shown in (17).

(17) A negative effect of token frequency (Pearson's r)

	Type frequency alone	Weighted by (log) token frequency
GCM	.743	.730
MGL	.767	.742

We see that the overall effect of token frequency weighting is quite small. The reason for this is that most words in the average corpus (and presumably also the average lexicon) have very low frequency (“Zipf’s Law”). As a result, weighting by token frequency influences just a small number of high-frequency words. Therefore, weighting by token frequency has relatively little effect, unless the target word happens to be very similar to an existing high-frequency word. It should be noted that these particular experimental items were not constructed for the purpose of dissociating type and token frequency, and ultimately the fairest test would be based on items that diverge more in their predictions. Nonetheless, the trend is clear across both models: to the extent that token frequency makes a difference, it is harmful in modeling speaker intuitions about the strength of the diphthongization pattern.

Like variegated similarity, high token frequency is a type of information that speakers could logically make use of in deciding whether or not to generalize a pattern to novel items. The fact that they apparently do not do so requires a formal model that is similarly restricted. As noted above, it is certainly possible to construct exemplar models that ignore token frequency; the amount and nature of frequency weighting is an independent parameter in the GCM that can be turned off completely, and Bybee (1995) explicitly defines schema strength in terms of type frequency. Conceptually, however, part of the appeal of exemplar models is that they rely on no special mechanisms except activating memory traces—a mechanism that intrinsically leads to token frequency effects (Bybee 2006). Insensitivity to token frequency follows quite naturally from a grammar of rules, however, since rules encode information that has been abstracted away from the particular exemplars that led to their creation. A rule-based account of analogy therefore involves no particular expectation that token frequency should play a role, and indeed is naturally restricted not to have access to information about token frequency.

9.4 The directionality of analogical inference

In the preceding sections, we have seen that an adequate model of analogical inference must be able to identify properties that are consistently associated with membership in a particular class, and must ensure that the association holds for sufficiently many different word types. Models that can find support for inferences in other ways, such as unstructured similarity or high token frequency, end up overestimating the goodness of many outcomes. A model without these abilities is more constrained, and has the advantage that it can more narrowly predict which analogical inferences speakers actually make. In this section I discuss one final restriction, concerning the direction of analogical inference.

Logically, statements about the relation between one form and another could be made in either direction. For example, statements about the correspondence of stressed and unstressed root allomorphs could relate either form to the other, symmetrically or asymmetrically, as in (18). This means that in principle, analogical inferences could proceed in multiple directions, both from stressless to stressed (e.g., *rentár:rénta* :: *sentár:*sénta*) and stressed to stressless (e.g., *siénta:sentár* :: *oriénta:*orentár*).

- (18) Some logically possible directions of influence (solid and doubled lines represent progressively greater pattern strength)

a. Stressed Unstressed	b. Stressed Unstressed	c. Stressed Unstressed
$\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} e$ $j\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} je$	$\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} e$ $j\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} je$	$\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} e$ $j\acute{e} \begin{array}{c} \longleftarrow \\ \rightleftarrows \\ \longrightarrow \end{array} je$

What we observe, however, is a striking restriction: both in historical change (Penny 2002; Morris 2005) and child errors (Clahsen, Aveledo, and Roca 2002), there is an overwhelming (or even exclusive) tendency for analogical rebuilding of stressed forms (i.e., *rentár:rénta* :: *sentár:*sénta*), consistent with (18b).⁸ A typical example from the Spanish portion of CHILDES is given in (19).

- (19) Overgeneralization of stressed mid vowels (Jorge, age 6; 1)
 y estonces **volo* a la pastelería
 and then (= *entonces*) fly-1sg (= *vuelo*) to the pastry shop
 ‘... and then I fly to the pastry shop’

⁸ Rebuilding stressless forms to include diphthongs has been reported in some dialects of Spanish (Judeo-Spanish, New Mexico Spanish). This data should be treated with care, however, since the morphology of these dialects also differs in more radical ways from literary Spanish. A similar effect is also reported in the experimental results of Bybee and Pardo (1981), but my preliminary attempts to replicate this finding have so far been unsuccessful.

Remarkably, the converse error (e.g., infinitive **vuelar* instead of *volar*) never occurs, and children also apparently never substitute mid vowels for nonalternating diphthongs (e.g., *él *frecónta* ‘he frequents’ instead of *frecuénta*). Similarly asymmetric error patterns have also been observed for Greek (Kazazis 1969), German (Clahsen, Aveledo, and Roca 2002) and Korean (Kang 2006), and appear to be the norm among children acquiring languages with morphophonological alternations. An explanatory model of analogy must be able to capture and ideally even predict such asymmetries.

Characterizing the direction of analogy has been a longstanding preoccupation in the historical linguistics literature, and numerous tendencies have been observed (Kurylowicz 1947; Mańczak 1980; Bybee 1985, and many others). The Spanish case seems atypical in several respects. It has sometimes been claimed that more frequent paradigm members are more influential (Mańczak 1980; Bybee 1985). In Spanish, the most frequent paradigm members (3sg, 1sg, 2sg) are all stressed, which should favor a stressed → stressless direction of influence. What we observe, however, is that the more frequent stressed forms are rebuilt on the basis of the less frequent stressless forms, counter to the more usual trend. Furthermore, it is often the case that the most influential forms are also less marked (in some intuitive sense of morphosyntactic markedness). What we see in Spanish, however, is that the 3sg, which is almost universally agreed to be the least marked combination of person and number features, is rebuilt on the basis of non-singular, non-third-person forms. Furthermore, diphthongs appear in the majority of present-tense indicative forms (the 1sg, 2sg, 3sg, and 3pl = 4 out of 6), yet reanalysis is done on the basis of the minority stressless forms. In short, the direction of influence that prevails in Spanish does not appear to follow from any general principle of frequency or markedness.

Albright (2002a, b) proposes that speakers generalize in some directions and not others because of a restriction on how paradigm structure is encoded. In particular, it is proposed that paradigms have an intrinsically asymmetrical organization in which certain forms are designated as “basic” and the remaining forms are derived from them by grammatical rules. For example, the error data suggests that in Spanish, a stressless form of the root (as found in the infinitive, 1pl, or 2pl) is taken as basic, and stressed forms are predicted—sometimes incorrectly—on the basis of a stressless form. The challenge is to understand why Spanish speakers choose this particular direction, and why paradigm organization may differ from language to language.

One principle of paradigm organization, explored also by Finkel and Stump (this volume), is PREDICTABILITY: a form is basic (\approx a PRINCIPAL PART) if it contains enough information to predict other forms in the

paradigm. As Finkel and Stump point out, there are many ways in which paradigms could be organized around predictive forms, depending on how many basic forms we are allowed to refer to, whether paradigm structure may differ from class to class, and so on. Many paradigm-based theories of morphology designate specific forms as “reference forms” in one way or another, and use these forms as the basis of computation for the remaining forms in the paradigm (Wurzel 1989; Stump 2001; J. P. Blevins 2006). Albright (2002a) adopts a particularly restrictive hypothesis: paradigm structure is the same (static) across all lexical items, and each form in the paradigm is based on just one other base form. The task of the learner is to find the base forms that permit the most accurate mappings, while still obeying this restriction.

The base identification algorithm, in brief, works as follows: the learner starts with a small batch of initial input data, consisting of paradigmatically related forms (1sg, 2sg, 3sg, etc.). Each one of these forms is considered as a potential base form, and the minimal generalization learner is used to find sets of rules that derive the remaining forms in the grammar. The result is a set of competing organizations, shown in Figure 9.3. In the usual case, at least some parts of the paradigm suffer from phonological or morphological neutralizations, with the result that not every form is equally successful at predicting the remainder of the paradigm. In these cases, some of the competing grammars will be less certain or accurate than others. The learner compares the candidate organizations to determine which form is associated with the most accurate rules, and this is chosen as the base for the remainder of the paradigm. This process may also be run recursively among the derived forms, to establish additional intermediate bases. (See Albright 2002a, b for details.)

When this procedure is run on an input of Spanish present-tense verb paradigms, the organization in Figure 9.4 results. Crucially, due to the restriction that each form be based on exactly one other base form, the model allows only five possible directions of inference (out of $5 \times 6 = 30$ logical possible pairwise relations). Some of these relations, such as infinitive

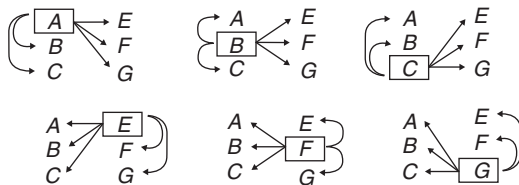


FIGURE 9.3 Candidate grammars, using asymmetrical mappings from a single base

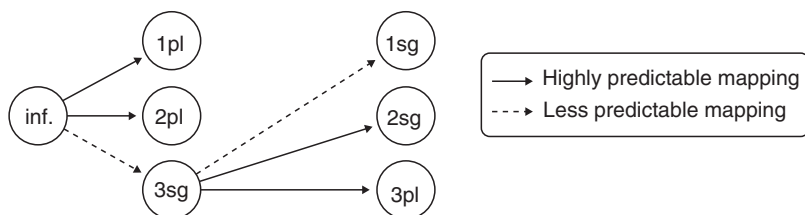


FIGURE 9.4 Predicted organization of Spanish present-tense paradigms

→ 1pl or 3sg → 2sg, are virtually 100% predictable, and leave no room for error. The greatest opportunities for analogical errors involve the mapping from stressless to stressed forms (here, infinitive → 3sg), and to the 1sg in particular. In fact, both of these mappings correspond to attested child errors:

- (20) Stem errors among children acquiring Spanish (Clahsen, Aveledo, and Roca 2002)
- a. Stressed stem replaced by stressless stem:
 - **volo* for *vuelo* ‘fly-1sg’; **juga* for *juega* ‘play-3sg’; **tene* for *tiene* ‘have-3sg’; **teno* for *tengo* ‘have-1sg’
 - b. Irregular 1sg replaced by stem from 3sg:
 - **tieno* for *tengo* ‘have-1sg’; **sabo* for *sé* ‘I know’; **conozo* for *conozco* ‘I know’; **parezo* for *parezco* ‘I appear’; **salo* for *salgo* ‘I leave’; **oyo* for *oigo* ‘I hear’

Although this analysis is somewhat skeletal and leaves many broader questions about paradigm structure unanswered,⁹ it highlights some of the virtues of a rule-based model of analogy. In particular, grammatical formalisms place strong restrictions on possible analogical inferences by dictating which forms may be effected, which patterns can be extended, and so on. Naturally, the strength and nature of these restrictions may vary considerably depending on the formalism; I have argued here in favor of a grammar of probabilistic context-sensitive rules that asymmetrically relate forms in the

⁹ In particular, it is natural to wonder how such a restrictive model could cope with systems that involve significantly more ambiguity—i.e., systems that motivate multiple principal parts in Finkel and Stump’s terms. It is important to keep in mind that nothing in the current model precludes the possibility that at a given point in time, languages may exhibit patterns that may be characterized as symmetrical predictability relations. A prediction of the asymmetrical model, however, is that learners will learn implications in just one direction, and that analogical generalizations should therefore go primarily in one direction. One type of data that is often telling in this regard is the relative size and frequency of the inflectional classes involved. Frequently classes that can be distinguished only in derived (nonbasic) forms are small and consist of words with high token frequency, which may be correlated with their status as memorized exceptions rather than as grammatically principled forms.

paradigm, but other formalisms are possible. The advantage of such a restrictive model is that it makes very specific and testable predictions about possible errors, and presumably also eventual historical changes. In the cases examined, these predictions appear to be substantially correct.

9.5 Conclusion

The results in the preceding sections have a common theme: in each case, the data of Spanish contains patterns that might logically lead to analogical inferences, yet speakers appear not to generalize them to novel or unknown items. I have argued that this reveals fundamental restrictions on how speakers learn to encode linguistic knowledge, which make these patterns either inaccessible or unimpressive. Furthermore, I have shown that a model based on probabilistic context-sensitive rules is well suited to capturing these restrictions. First, it limits the type of similarity relations that are relevant in supporting analogy: they must be “structured” in the sense that supporting analogs must all share a set of properties that are reliably correlated with class membership. As shown in Section 9.2, speakers, too, appear to obey this restriction, and models that lack such structure overpredict the goodness of many logically possible inferences. In addition, attributing analogy to a grammar of rules leads us to expect that generalizations should be based on high type frequency of similar words, and that token frequency should be irrelevant; in Section 9.3, we saw that this, too, appears to be correct. Finally, rewrite rules are an intrinsically directional formalism ($A \rightarrow B$), corresponding to the idea that inference proceeds in some directions but not others. In Section 9.4, I argued that a model of paradigm structure based on predictability relations between related forms can predict which directions speakers actually choose, in a way that appears to line up well with data from child errors and historical change. In each case, the payoff of the more restrictive formalism is clear: it provides an account for why some errors occur and some do not, providing a more explanatory model of how speakers carry out analogy in morphophonological systems.

The examples discussed here are also intended to highlight some virtues of computationally implemented models of analogy. At the most basic level, the models facilitate a quantitative assessment of the relative contribution of different types of analogical reasoning, by allowing us to compare directly the predictions of models with and without a particular capacity. Such comparisons are potentially quite important in an area where it is easy to posit many potentially relevant factors (high token frequency, semantic

effects, phonetic factors, etc.), but difficult to establish their explanatory value. Equally important, though, is the role that modeling may play in shaping and refining theoretical distinctions. An example of this was seen in Section 9.2.4, in which comparison of the two models required a more careful definition of the concept of structured similarity, and testing the distinction was only possible by interpreting one model with respect to the other. We are only beginning to develop the analytical tools needed to construct theoretical arguments from such modeling results. I hope to have shown, however, that computational modeling can play a role not only in testing, but also in developing theories of what constitutes a possible analogy.

Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation

*Petar Milin, Victor Kuperman, Aleksandar Kostić,
and R. Harald Baayen*

10.1 Introduction

Syntagmatically oriented theories of word structure have inspired most of the experimental work on morphological processing. The way inflection is modeled by Levelt *et al.* (1999), for instance, comes close to the theory of distributed morphology proposed by Halle and Marantz (1993). In Levelt's model of speech production, nodes at the lemma stratum (what would be the lexeme stratum in the terminology of Aronoff (1994)) are marked for features such as tense, aspect, number, and person. For a given set of feature values, a node at the form stratum will be activated, e.g., *-ed* for the past tense in English. Paradigmatic relations do not have a place in this model; in fact, it is a design feature of the model that paradigmatic relations at the level of word forms are predicted to be irrelevant.

A syntagmatic bias is also visible in the comprehension model proposed by Schreuder and Baayen (1995). In this model, there is no principled difference between stems or words on the one hand, and affixes (whether inflectional or derivational) on the other hand. In their three-layered network, with access units, lemma units, and semantic and syntactic feature units, the organization of nodes within a layer is arbitrary. Paradigmatic relations do not play a role, they are simply deemed to be irrelevant. The same holds for the dual mechanism model of Pinker (1991, 1999).

In this chapter, we present a line of research that departs from the syntagmatic orientation of mainstream experimental psycholinguistics, and that is close in spirit to word and paradigm morphology (WPM, Hockett 1954; Matthews 1974; Anderson 1992; Aronoff 1994; Beard 1995; J. P. Blevins 2003, 2006*b*). WPM questions the morphemic status of lexical formatives, and assumes that words (both simple and complex) are the basic units in the lexicon. Furthermore, in WPM, inflected words are organized into paradigms, which are further organized into inflectional classes.¹

From a processing perspective, the central tenets of WPM imply, first, that complex words, including regular inflected words, leave traces in long-term lexical memory, and second, that the processing of a given word is codetermined by paradigmatically related words.

A central diagnostic for the presence of memory traces in long-term memory has been the word frequency effect. A higher frequency of use allows for shorter processing latencies in both visual and auditory comprehension (cf. Baayen *et al.* 2003*a*; New *et al.* 2004; Baayen *et al.* 2006, etc.), and lower rates of speech errors in production (Stemberger and MacWhinney 1986). The effect of word frequency tends to be stronger for irregular complex words than for regular complex words, and stronger for derived words than for inflected words. But even for regular inflected words, the effect of prior experience clearly emerges (Baayen *et al.* 2008*c*), contrary to the claims of the dual mechanism model. The ubiquitous effect of word frequency shows that large numbers of complex words are indeed available in the (mental) lexicon, as claimed by WPM.

The focus of this chapter is on the second central processing consequence of WPM, namely, that paradigmatic organization should codetermine lexical processing. For derivational morphology, work on the morphological family size effect (see, e.g., Moscoso del Prado Martín *et al.* 2004) has clarified that processing of a given word is codetermined by other morphologically related words. This constitutes evidence for paradigmatic organization in the mental lexicon. However, morphological families are very heterogeneous, and do not readily allow words to be grouped into higher-order sets similar to inflectional classes. Therefore, the morphological family size effect provides at best circumstantial evidence for the central ideas of WPM.

In the remainder of this chapter, we first review a series of recent experimental studies which explore the role of paradigmatic structure specifically

¹ In what follows, we will use the term *inflectional paradigm* to refer to the set of inflected variants of a given lexeme, and the term *inflectional class* to refer to a set of lexemes that use the same set of exponents in their inflectional paradigms.

for inflected words. We then present new experimental results showing how the principles that structure inflectional paradigms can be generalized to subsets of derived words.

The approach to morphological organization and morphological processing that we describe in this chapter departs significantly from both theoretical morphology and mainstream of experimental psycholinguistics in that it applies central concepts from information theory to lexical processing. The greater the amount of information carried by an event (e.g., a word's inflected variant, an exponent, or an inflectional class), the smaller the probability of that event, and the greater the corresponding processing costs (see, for a similar approach to syntax, Levy 2008). We believe that information theory offers exactly the right tools for studying the processing consequences of paradigmatic relations. Furthermore, we do believe that the concepts of information science provide us with excellent tools to probe the *functional organization of the mental lexicon*, but we shall remain agnostic about how paradigmatic structures are implemented in the brain.

We begin this chapter with an introduction to a number of central concepts from information theory and illustrate how these concepts can be applied to the different levels of paradigmatic organization in the mental lexicon. We then focus on three key issues: (i) the processing cost of an exponent given its inflectional class, (ii) the processing cost associated with inflectional paradigms and inflectional classes, and (iii) the processing cost that arises when the probabilistic distributional properties of paradigms and classes diverge.

In what follows, we first provide a comprehensive review of previous experimental findings that use information-theoretic measures of lexical connectivity. We then present some new results that provide further empirical support for the relevance of paradigmatic organization for lexical processing, and for the importance of information-theoretic measures for gauging the processing consequences of paradigmatic structure. As we proceed through our discussion of the empirical evidence, it will become increasingly clear that there is a remarkable convergence between the psycholinguistic evidence and WPM.

Some of the key findings of the general approach to the (mental) lexicon outlined in this chapter can be summarized as follows:

1. Lexemes and their inflected variants are organized hierarchically. One can envision this organization as a higher layer of lexemes grouped into morphological families, and a lower level of inflected variants, which enter into paradigmatic relations within a given lexeme.

2. Inflected variants of any given lexeme are organized into paradigms, and all lexemes that form their paradigms in the same way define an inflectional class. Empirical evidence suggests that the degree to which the inflectional paradigm of a given lexeme diverges from its inflectional class affects cognitive processing over and above other relevant factors: the greater the divergence, the more costly the processing.
3. Results which will be presented here for the first time show that the processing of English derivatives can be seen as analogical. During lexical processing, a given derivative is compared with its base word, and pitted against the generalized knowledge about the relationship between all derivatives of the same type and their corresponding base words.
4. The *family size effect*, which is known to be a semantic effect, probably represents the joint effect of both semantic similarity and morphological paradigmatic structure.

10.2 Central concepts from information theory

A fundamental insight of information theory is that the amount of information I carried by (linguistic) unit u can be defined as the negative binary logarithm of its probability:

$$I_u = -\log_2 \Pr(u). \quad (1)$$

Consider someone in the tip-of-the-tongue state saying *the eh eh eh eh eh eh key*. The word *eh* has the greatest probability, $6/8$, and is least informative. Its amount of information is $-\log_2(6/8) = 0.415$ bits. The words *the* and *key* have a probability of $1/8$ and the amount of information they carry is 3 bits. In what follows, we assume that lexical units that have a higher information load are more costly to access in long-term memory. Hence, we expect processing costs to be proportional to the amount of information. This is exactly what the word frequency effect tells us: higher-frequency words, which have lower information loads, are processed faster than low-frequency, high-information words.

We estimate probabilities from relative frequencies. By way of illustration, consider the inflected variants of the Serbian feminine noun *planina* ('mountain'). Serbian nouns have six cases and two numbers. Due to syncretism, the twelve combinations of case and number are represented by only six distinct inflected variants. These inflected variants are listed in column 1 of the upper part of Table 10.1. The second column lists the frequencies of these inflected variants in a two-million word corpus of written Serbian.

TABLE 10.1 Inflected nouns in Serbian. The upper part of the table shows inflected variants for the feminine noun *planina* ('mountain'), the lower part shows the inflected variants of the masculine noun *prostor* ('space'). Columns present frequencies and relative frequencies of the respective inflectional paradigm and the class to which it belongs.

feminine nouns						
Inflected variant	Inflected variant frequency	Inflected variant relative frequency	Information of inflected variant	Exponent frequency	Exponent relative frequency	Information of exponent
	$F(w_e)$	$\text{Pr}_\pi(w_e)$	I_{w_e}	$F(e)$	$\text{Pr}_\pi(e)$	I_e
planin- <i>a</i>	169	0.31	1.69	18715	0.26	1.94
planin- <i>u</i>	48	0.09	3.47	9918	0.14	2.84
planin- <i>e</i>	191	0.35	1.51	27803	0.39	1.36
planin- <i>i</i>	88	0.16	2.64	7072	0.1	3.32
planin- <i>om</i>	30	0.05	4.32	4265	0.06	4.06
planin- <i>ama</i>	26	0.05	4.32	4409	0.06	4.06
masculine nouns						
Inflected variant	Inflected variant frequency	Inflected variant relative frequency	Information of inflected variant	Exponent frequency	Exponent relative frequency	Information of exponent
	$F(w_e)$	$\text{Pr}_\pi(w_e)$	I_{w_e}	$F(e)$	$\text{Pr}_\pi(e)$	I_e
prostor- <i>φ</i>	153	0.38	1.40	25399	0.35	1.51
prostor- <i>a</i>	69	0.17	2.56	18523	0.26	1.94
prostor- <i>u</i>	67	0.17	2.56	8409	0.12	3.06
prostor- <i>om</i>	15	0.04	4.64	3688	0.05	4.32
prostor- <i>e</i>	48	0.12	3.06	5634	0.08	3.64
prostor- <i>i</i>	23	0.06	4.06	6772	0.09	3.47
prostor- <i>ima</i>	23	0.06	4.06	3169	0.04	4.64

We consider two complementary ways of estimating probabilities from frequencies. The probabilities listed in the third column of Table 10.1 are obtained by normalizing the frequency counts with respect to a lexeme's inflectional paradigm (column three). More specifically, the probability $\text{Pr}_\pi(w_e)^2$ of an inflected variant w_e of lexeme w is estimated in this table as

² Here and in what follows we use Pr_π to denote probabilities defined with respect to paradigmatic sets.

its form-specific frequency F (henceforth *word frequency*) of occurrence, normalized for the sum of the frequencies of all the distinct inflected variants of its lexeme, henceforth *stem frequency*:

$$\Pr_{\pi}(w_e) = \frac{F(w_e)}{\sum_e F(w_e)}. \quad (2)$$

The corresponding amounts of information, obtained by applying (1), are listed in column four. Table 10.1 also lists the frequencies of the six exponents (column 5), calculated by summing the word frequencies of all forms in the corpus with these exponents. The probabilities listed for these exponents (column six) are obtained by normalizing with respect to the summed frequencies of these exponents:

$$\Pr_{\pi}(e) = \frac{F(e)}{\sum_e F(w_e)}. \quad (3)$$

The corresponding amount of information is listed in column seven.

The second way in which we can estimate probabilities is by normalizing with respect to the number of tokens N in the corpus. The probability of a lexeme w is then estimated as the sum of the frequencies of its inflected variants, divided by N :

$$\Pr_N(w) = \frac{F(w)}{N} = \frac{\sum_e F(w_e)}{N}. \quad (4)$$

In this approach, the probability of an inflected variant can be construed as the joint probability of its lexeme w and its exponent:

$$\begin{aligned} \Pr_N(w_e) &= \Pr(w, e) \\ &= \Pr(e, w) \\ &= \frac{F(w_e)}{N}. \end{aligned} \quad (5)$$

Likewise, the probability $\Pr(e)$ of an exponent (e.g., $-a$ for nominative singular and genitive plural in Serbian feminine nouns) can be quantified as the relative frequency of occurrence of e in the corpus:

$$\Pr_N(e) = \frac{F(e)}{N}. \quad (6)$$

The probabilities considered thus far are unconditional, a priori, decontextualized probabilities. As exponents appear in the context of stems, we

need to consider the conditional probability of an exponent given its lexeme, $\Pr(e|w)$. Using Bayes' theorem, we rewrite this probability as:

$$\begin{aligned}\Pr_N(e|w) &= \frac{\Pr_N(e,w)}{\Pr_N(w)} \\ &= \frac{F(w_e)}{N} \frac{N}{F(w)} \\ &= \frac{F(w_e)}{F(w)} \\ &= \Pr_\pi(w_e).\end{aligned}\tag{7}$$

Likewise, the conditional probability of the lemma given the exponent is defined as:

$$\begin{aligned}\Pr_N(w|e) &= \frac{\Pr_N(w,e)}{\Pr_N(e)} \\ &= \frac{F(w_e)}{N} \frac{N}{F(e)} \\ &= \frac{F(w_e)}{F(e)}.\end{aligned}\tag{8}$$

For each lexical probability we can compute the corresponding amount of information. We allow for the possibility that each source of information may have its own distinct effect on lexical processing by means of positive weights ω_{1-5} :

$$\begin{aligned}I_{w_e} &= -\omega_1 \log_2 F(w_e) + \omega_1 \log_2 N \\ I_w &= -\omega_2 \log_2 F(w) + \omega_2 \log_2 N \\ I_e &= -\omega_3 \log_2 F(e) + \omega_3 \log_2 N \\ I_{e|w} &= -\omega_4 \log_2 F(w_e) + \omega_4 \log_2 F(w) \\ I_{w|e} &= -\omega_5 \log_2 F(w_e) + \omega_5 \log_2 F(e).\end{aligned}\tag{9}$$

We assume that the cost of retrieving lexical information from long-term memory is proportional to the amount of information retrieved. Hence the cost of processing an inflected word w_e is proportional to at least the amounts of information in (9). More formally, we can express this processing cost (measured experimentally as a reaction time RT) as a linear function:

$$\begin{aligned}
 RT &\propto I_{w_e} + I_w + I_e + I_{e|w} + I_{w|e} \\
 &= (\omega_1 + \omega_2 + \omega_3) \log_2 N - (\omega_1 + \omega_4 + \omega_5) \log_2 F(w_e) \\
 &\quad - (\omega_2 - \omega_4) \log_2 F(w) - (\omega_3 - \omega_5) \log_2 F(e). \tag{10}
 \end{aligned}$$

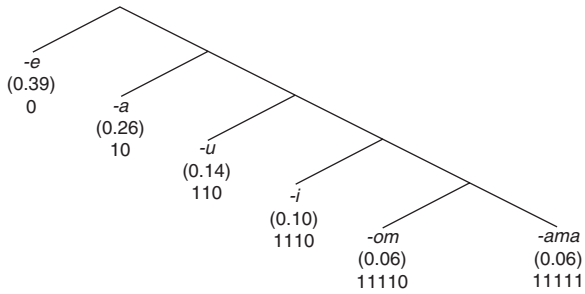
There are several predictions for the effects of lexical probabilities on lexical processing that follow directly from (10). First, word frequency $F(w_e)$ will always elicit a facilitatory effect, as all its coefficients have a negative sign in (10). Second, stem frequency $F(w)$ may either facilitate or inhibit processing, depending on the relative strengths of the coefficients ω_2 and ω_4 . These two coefficients balance the importance of a word's probability as such (see the second equation in (9)), and its importance as the domain on which the probabilities of its inflectional variants are conditioned (see the fourth equation in (9)). Third, the frequency of the exponent can also either speed up or hinder processing depending on the values of ω_3 and ω_5 . These two weights balance the importance of an exponent's probability as such (see the first equation in (9)) and the exponent as the domain on which the probability of inflected forms with that exponent are conditioned (see the fifth equation in (9)). The first two predictions are supported by the large-scale regression studies reported by Baayen *et al.* (2008c) and Kuperman *et al.* (2008).

We now proceed from basic lexical probabilities that operate at the level of individual inflected words to the quantification of the information carried by inflectional paradigms and inflectional classes. The paradigm of a given lexeme can be associated with a distribution of probabilities $\{\text{Pr}_\pi(w_e)\}$. For *planina* in Table 10.1, this probability distribution is given in column three. The amount of information carried by its paradigm as a whole is given by the *entropy* of the paradigm's probability distribution:

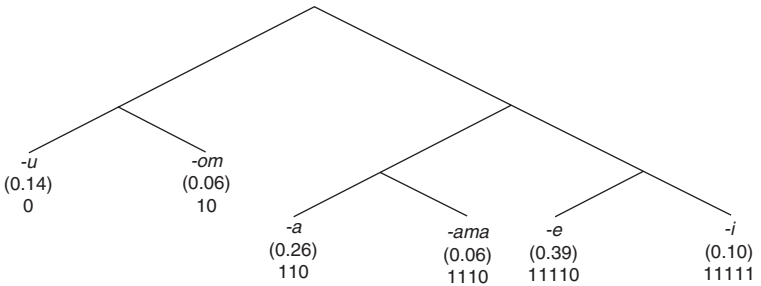
$$H = - \sum_e \text{Pr}_\pi(w_e) \log_2 (\text{Pr}_\pi(w_e)). \tag{11}$$

Formally, H is the expected (weighted average) amount of information in a paradigm. The entropy increases with the number of members of the paradigm. It also increases when the probabilities of the members are more similar. For a given number of members, the entropy is maximal when all probabilities are the same. H also represents the average number of binary decisions required to identify a member of the paradigm, i.e., to reduce all uncertainty about which member of the paradigm is at issue, provided that the paradigm is represented by an optimal binary coding. We illustrate the concept of optimal coding in Figure 10.1 using as an example the inflectional class of regular feminine nouns in Serbian.

BIT = 2.33



BIT = 2.83



BIT = 4.29

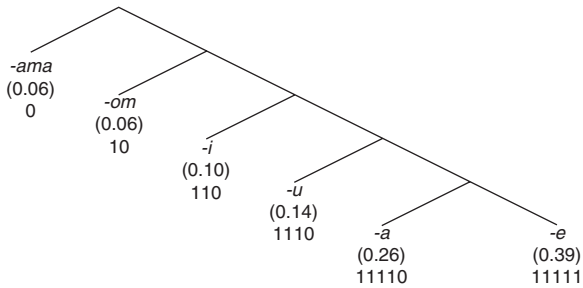


FIGURE 10.1 Optimal and nonoptimal binary coding schemes for the inflectional class of regular feminine nouns in Serbian.

The upper panel of Figure 10.1 shows an optimal binary coding scheme, in which the most probable exponent ($-e$, $\Pr_{\pi} = 0.39$) occupies the highest leaf node in the tree. The lower the probability of the other exponents, the lower in the tree they are located. Thus, the exponents with the lowest probabilities in the inflectional class, $-om$ ($\Pr_{\pi} = 0.06$) and $-ama$ ($\Pr_{\pi} = 0.06$) are found at the lowest leaf nodes. The second panel of Figure 10.1 represents another possible coding, which is suboptimal in that some exponents with relatively

high probabilities are located below lower-probability exponents in the tree. Finally, the third panel shows the least optimal coding, in which the less probable the exponent is, the *higher* it is positioned in the tree. The average number of binary decisions (the number of bits) required to identify a given paradigm member, i.e., to reach the paradigm member's leaf node when starting at the root node of the tree, is the sum of the products of the number of steps and the members' probabilities. This average is never greater than the entropy of the paradigm $H + 1$ (Ross 1988). For the upper panel of Figure 10.1, the average number of binary decisions is 2.33 bits, for the coding in the second panel, it is 2.83, and for the worst coding in the third panel, it is 4.29. In Section 10.4 we will review experimental studies showing that paradigmatic entropies codetermine lexical processing.

Thus far, we have considered probabilities and the corresponding entropy at the level of the inflectional class of regular feminine nouns in Serbian. However, the probability distribution of the inflected variants of a given lexeme may differ substantially from the probability distribution of the exponents at the level of the inflectional class. As a consequence, the corresponding entropies may differ substantially from each other as well. The extent to which these probability distributions differ is quantified by the relative entropy, also known as Kullback-Leibler divergence. Consider again the Serbian feminine noun *planina* 'mountain' and its inflectional class as shown in Table 10.1. The third column lists the estimated probabilities for the paradigm, and the sixth column lists the probability distribution of the class. Let P denote the probability distribution of the paradigm, and Q the probability distribution of the inflectional class. The relative entropy can now be introduced as:

$$D(P \parallel Q) = \sum_e \Pr_\pi(w_e) \log_2 \frac{\Pr_\pi(w_e)}{\Pr_\pi(e)}. \quad (12)$$

Relative entropy is also known as *information gain*,

$$\begin{aligned} D(P \parallel Q) &= IG(\Pr_\pi(e|w) \parallel \Pr_\pi(e|c)) \\ &= \sum_e \Pr_\pi(e|w) \log_2 \frac{\Pr_\pi(e|w)}{\Pr_\pi(e|c)} \\ &= \sum_e \Pr_\pi(w_e) \log_2 \frac{\Pr_\pi(w_e)}{\Pr_\pi(e)}, \end{aligned} \quad (13)$$

as it measures the reduction in our uncertainty about the exponent (e) when going from the situation in which we only know its inflectional class (c) to the situation in which we also know the lexeme (w). For *planina*, $H = 2.22$,

and $D(P \parallel Q) = 0.05$. For the masculine noun *prostor* listed in the lower half of Table 10.1, $H = 2.42$ and $D(P \parallel Q) = 0.07$. In both cases, the two distributions are fairly similar, so the relative entropies (*RE*) are small. There is little that the knowledge of *planina* adds to what we already new about regular feminine nouns. If we approximate the probability distribution of *planina* with the probability distribution of its class, we are doing quite well. In Section 10.4.2 we review a recent study demonstrating that *RE* is yet another information-theoretic predictor of lexical processing costs.

We will now review a series of studies that illustrate how these information theoretic concepts help us to understand paradigmatic organization in the mental lexicon. Section 10.3 addresses the question of how the probability of an exponent given its inflectional class is reflected in measures of lexical processing costs. Section 10.4 reviews studies that make use of entropy and relative entropy to gauge lexical processing and paradigmatic organization. Finally, in Section 10.5 we present new experimental results showing how concepts from information theory that proved useful for understanding inflection can help understanding derivation.

10.3 The structure of inflectional classes

The consequence of the amount of information carried by an exponent for lexical processing has been explored in a series of experimental studies on Serbian (Kostić 1991, 1995; Kostić *et al.* 2003). A starting point for this line of research is the amount of information carried by an exponent,

$$I_e = -\log_2 \text{Pr}_\pi(e),$$

where Pr_π is estimated over all exponents within a class π . Kostić and colleagues noted that exponents are not equal with respect to their functional load. Some exponents (given their inflectional class) express only a few functions and meanings, others express many. Table 10.2 lists the functions and meanings for the exponents of the masculine and regular feminine inflectional class of Serbian. The count of numbers of functions and meanings for a given exponent were taken from an independent comprehensive lexicological survey of Serbian (see also the appendix of Kostić *et al.* 2003, for a shortlist of functions and meanings). Instead of using just the flat corpus-based relative frequencies, Kostić and colleagues propose to weight these probabilities for their functions and meanings. Let R_e denote the number of functions and meanings carried by exponent e . Then the weighted amount of information I'_e can be expressed as follows:

TABLE 10.2 Exponents, case and number, frequency of the exponent, number of functions and meanings of the exponents, and amount of information carried by the exponents, for masculine nouns (upper table) and regular feminine nouns (lower table).

masculine nouns				
Exponent	Case and Number	Frequency	Functions and Meanings	Information
ϕ	nom sg	12.83	3	0.434
a	gen sg/acc sg /gen pl	18.01	109	5.128
u	dat sg /loc sg	4.64	43	5.744
om	ins sg	1.90	32	6.608
e	acc pl	2.21	58	7.243
i	nom pl	3.33	3	2.381
ima	dat pl/loc pl/ins pl	1.49	75	8.186
feminine nouns				
Exponent	Case and Number	Frequency	Functions and Meanings	Information
a	nom sg/gen pl	12.06	54	1.464
u	acc sg	5.48	58	2.705
e	gen sg /nom pl/acc pl	14.20	112	2.280
i	dat sg /loc sg	3.80	43	2.803
om	ins sg	1.94	32	3.346
ama	dat pl/loc pl/ins pl	1.69	75	4.773

$$I'_e = -\log_2 \left(\frac{\Pr_\pi(e)/R_e}{\sum_e \Pr_\pi(e)/R_e} \right) \quad (14)$$

The ratio $(\Pr_\pi(e)/R_e)$ gives us the average probability per syntactic function/meaning for a given exponent. In order to take the other exponents within the inflectional class into account, this ratio is weighted by the sum of the ratios for each of the exponents (see, e.g., Luce 1959). The resulting proportion is log-transformed to obtain the corresponding amount of information in bits. The partial effects of probability on the one hand, and the number of functions and meanings on the other, are shown in Figure 10.2. The weighted information is predicted to decrease with probability, and to increase with the number of functions and meanings. Table 10.2 lists I'_e for each of the exponents of the masculine and regular feminine inflectional classes.

To assess the predictivity of I'_e , Kostić *et al.* (2003) and Kostić (2008) calculated the mean lexical decision latency for each exponent in a given inflectional class, and investigated whether these mean latencies can be

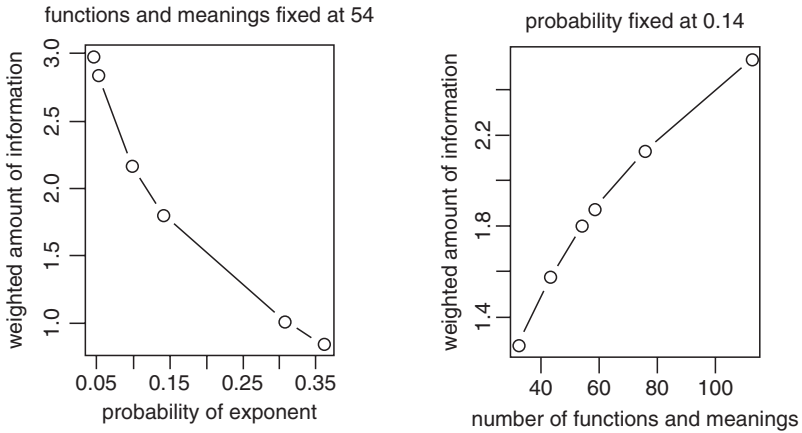


FIGURE 10.2 Partial effects of the probability of an exponent and its number of syntactic functions and meanings on the weighted amount of information I'_e .

predicted from the weighted amounts of information such as those listed in Table 10.2. The Pearson correlation between the mean latencies and the weighted information scores was highly significant for both masculine and feminine nouns ($R^2 = 0.88$ for masculine nouns, $R^2 = 0.98$ for regular feminine nouns and $R^2 = 0.99$ for irregular feminine nouns). Furthermore, when mean reaction time is regressed on the weighted information load, the slopes of the regression lines are positive. Exponents carrying a greater average amount of information are more difficult to process. In other words, these data show that the average processing cost of an exponent in its inflectional class is very well predicted from its frequency and its functional load as given by (14) and illustrated above in Figure 10.2.

The probabilities that we considered in these analyses were estimated by summing across all words with a given exponent in a given inflectional class. In this way, the information about the probabilities of the different exponents in the inflectional paradigms of specific words is lost. In order to address the possibility that word-specific probabilities of exponents also codetermine lexical processing, Kostić *et al.* (2003) first applied the same weighting scheme underlying (14) at the level of individual lexemes, giving a lexeme-specific weighted information I'_{w_e} :

$$I'_{w_e} = -\log_2 \left(\frac{\Pr_\pi(w_e)/R_e}{\sum_e \Pr_\pi(w_e)/R_e} \right). \quad (15)$$

Kostić *et al.* (2003) then constructed two sets of lexemes (henceforth Inflectional Groups) which contrasted maximally with respect to I'_{w_e} . For each of

the two inflectional groups, the average value of I'_{w_e} for each of the exponents was calculated. Regression analysis showed that these group-averaged amounts of information contributed independently to the model, over and above the general class-based information values I'_e . As before, larger values for the group-averaged amounts of information I'_{w_e} corresponded to longer mean lexical decision latencies.

It is useful to probe the lexeme-specific weighted information (15) with respect to how it relates to the frequency properties of the lexeme and its inflected variants, as well as to the functional ambiguities existing in inflectional paradigms and classes. First, consider a simple lower bound for (15):

$$\begin{aligned}
 I'_{w_e} &= -\log_2 \left(\frac{\Pr_{\pi}(w_e)/R_e}{\sum_e \Pr_{\pi}(w_e)/R_{w_e}} \right) \\
 &= -\log_2 \frac{\Pr_{\pi}(w_e)}{R_e} + \log_2 \sum_e \frac{\Pr_{\pi}(w_e)}{R_e} \\
 &\geq -\log_2 \Pr_{\pi}(w_e) + \log_2 R_e + \log_2 \prod_e \frac{\Pr_{\pi}(w_e)}{R_e} \\
 &\geq -\log_2 \Pr_{\pi}(w_e) + \log_2 R_e + \sum_e \log_2 \frac{\Pr_{\pi}(w_e)}{R_e} \\
 &\geq \log_2 R_e - \sum_e \log_2 R_e - \log_2 \Pr_{\pi}(w_e) + \sum_e \log_2 \Pr_{\pi} w_e. \quad (16)
 \end{aligned}$$

The third term is the amount of information carried by the inflected variant, I_{w_e} , see (2), and $\sum_j \log_2 \Pr_{\pi} w_j$ is a measure of the lexeme's stem frequency, evaluated by summing the log frequencies of its inflected variants rather than by summing the bare frequencies of its inflected variants. Consequently, at the level of the inflected variant, the amount of information (16) incorporates two well-known frequency effects that have been studied extensively in the processing literature. The word frequency effect ($-\log_2 \Pr_{\pi}(w_e)$) is facilitatory, as expected. Surprisingly, the stem frequency effect ($\sum_e \log_2 \Pr_{\pi} w_e$) is predicted to be inhibitory. However, both frequency effects are complemented by measures gauging ambiguity. Ambiguity of the given exponent is harmful, whereas ambiguity in the rest of the paradigm is facilitatory. Thus, the stem frequency effect emerges from this model as a composite effect with both an inhibitory and a facilitatory component. This may help explain why stem frequency effects are often much less robustly attested in experimental data (see, e.g., Baayen *et al.* 2008c) compared to word frequency effects.

In order to evaluate how well the lower bound given in (16) approximates the original measure given in (15), we examined the exponent frequency, the

TABLE 10.3 Mean reaction times in visual lexical decision (RT), exponent frequency, number of functions and meanings of the exponent (R), weighted amount of information (I'), and Inflectional Group (high versus low by-word amount of information) for the Exponents of the regular feminine declension class.

Exponent	Exponent frequency	R	I'	Inflectional Group	RT
<i>a</i>	12.06	3.99	1.46	high	674
<i>e</i>	14.20	4.72	2.28	high	687
<i>i</i>	3.80	3.76	2.80	high	685
<i>u</i>	5.48	4.06	2.71	high	693
<i>om</i>	1.94	3.47	3.35	high	718
<i>ama</i>	1.69	4.32	4.77	high	744
<i>a</i>	12.06	3.99	1.46	low	687
<i>e</i>	14.20	4.72	2.28	low	685
<i>i</i>	3.80	3.76	2.80	low	730
<i>u</i>	5.48	4.06	2.71	low	712
<i>om</i>	1.94	3.47	3.35	low	722
<i>ama</i>	1.69	4.32	4.77	low	746

group averages of the functions and meanings, the information values, and the mean reaction times for the two inflectional groups for regular feminine nouns, as listed in Table 10.3 (data from Kostić *et al.* 2003). Note that the terms in (16) represent the ambiguity of the exponent, the joint ambiguity of all exponents, the word frequency effect of the inflected variant, and the stem frequency effect of its lexeme.

For the data in Table 10.3, we first carried out a linear regression analysis with RT as dependent variable and I' and Inflectional Group as predictors. The R^2 for this model was 0.863. We then carried out a linear regression analysis, but now with the two measures that figure in the lower bound of the amount of information (16) as predictors: exponent frequency and the number of functions and meanings of the exponent R. The R^2 of this model was 0.830. Furthermore, the effect of the number of functions and meanings was inhibitory ($\hat{\beta} = 27.5, t(8) = 2.512, p = 0.0362$) and the effect of exponent frequency was facilitatory ($\hat{\beta} = -5.2, t(8) = -5.813, p = 0.0004$) as expected given (16). In other words, the two variables that according to (16) should capture a substantial proportion of the variance explained by the amount of information I' , indeed succeed in doing so: 0.830 is 96 percent of 0.863.

The lower bound estimate in (16) is a simplification of the full model I'_{w_c} defined by (15). Because the simplification allows us to separate the word and stem frequency effects, it clarifies that these two frequency effects are given the same overall weight. There is evidence, however, that stem frequency has a much more modest weight than word frequency (Baayen *et al.* 2008c), and

may even have a different functional form. This suggests that it may be preferable to rewrite (15) as:

$$I'_{w_e} = -\log_2 \left(\frac{\omega_1 \Pr_\pi(w_e)/R_e}{\omega_2 \sum_e \Pr_\pi(w_e)/R_e} \right), \quad (17)$$

with separate weights ω for numerator and denominator. On the other hand, at the level of a given class the lower bound estimate in (17) reduces to the exponent frequency and the overall class frequency. Some preliminary experimental evidence for the relevance of exponent frequency (in the simplified form of inflectional formative frequency) for English is available in Baayen *et al.* (2008c), along with evidence for frequency effects for derivational affixes. However, it is presently unclear how class frequency could be generalized and gauged with derivations. Inflectional classes are well contained and it is easy to count out their overall frequencies. Contrariwise, within and between derivational classes there are no clear partitions of the lexical space. While inflected words, in general, belong to only one inflectional class, any given base word may participate in several derivations. We shall address the issue of relations between base words and their derivatives in codetermining lexical processing in further detail in Section 10.5.

It is also useful to rewrite (14) along similar lines to what we did for (15). In this case, the lower bound for the amount of information can be written as the sum of two conditional probabilities. First, consider the probability of exponent e given its inflectional class c :

$$\begin{aligned} \Pr(e|c) &= \frac{\Pr(e,c)}{\Pr(c)} \\ &= \frac{\Pr(e)}{\Pr(c)}. \end{aligned}$$

(Note that the probability of an exponent is defined strictly with respect to its inflectional class. We never sum frequencies of exponents across inflectional classes.) The information corresponding to this conditional probability is

$$\begin{aligned} I_{e|c} &= -\log_2 \frac{\Pr(e)}{\Pr(c)} \\ &= -\log_2 \Pr(e) + \log_2 \Pr(c) \\ &= -\log_2 \Pr(e) + \log_2 \sum_j \Pr(e_j) \end{aligned}$$

$$\begin{aligned}
&\geq -\log_2 \Pr(e) + \log_2 \prod_j \Pr(e_j) \\
&\geq -\log_2 \Pr(e) + \sum_j \log_2 \Pr(e_j) \\
&= I'_{e|c}
\end{aligned} \tag{18}$$

Note that $I'_{e|c}$ is a lower bound of $I_{e|c}$.

Next, let R_e denote the number of functions and meanings of exponent e in class c , and let R_c denote the total count of functions and meanings within the class. The conditional probability of the functions and meanings of exponent e given its class c is

$$\begin{aligned}
\Pr(R_e|R_c) &= \frac{\Pr(R_e, R_c)}{\Pr(R_c)} \\
&= \frac{\Pr(R_e)}{\Pr(R_c)} \\
&= \frac{R_e}{R_c}
\end{aligned}$$

and the corresponding information is therefore

$$\begin{aligned}
I_{R_e|R_c} &= -\log_2 \frac{R_e}{R_c} \\
&= -\log_2 R_e + \log_2 R_c \\
&= -\log_2 R_e + \log_2 \sum_j R_j \\
&\leq -\log_2 R_e + \log_2 \prod_j R_j \\
&\leq -\log_2 R_e + \sum_j \log_2 R_j \\
&= I'_{R_e|R_c}
\end{aligned} \tag{19}$$

Here, $I'_{R_e|R_c}$ is an upper bound of $I_{R_e|R_c}$.

Taking into account that $I'_{e|c}$ is a lower bound of $I_{e|c}$ and that $I'_{R_e|R_c}$ is an upper bound of $I_{R_e|R_c}$, we can now approximate (14) as follows:

$$\begin{aligned}
I_{w_e} &\approx \log_2 R_e - \sum_j \log_2 R_j - \log_2 \Pr_{\pi} w_e + \sum_j \log_2 \Pr_{\pi} w_j \\
&\approx -I'_{R_e|R_c} + I'_{e|c}.
\end{aligned} \tag{20}$$

In other words, the amount of information as defined in (14) is related to the sum of two conditional probabilities: (i) the probability of the exponent given its class, and (ii) the probability of the ambiguity of the exponent given the ambiguity in its class. The partial effects of these two conditional probabilities are shown in Figure 10.3. As expected, the partial effects are very similar to those shown in Figure 10.2.

At this point, the question arises why $I'_{R_e|R_c}$ appears with a negative sign in (20). To answer this question, we need to consider exponents within their classes, and differentiate between the functions and meanings that an inflected form can have in the discourse. Consider the case in which $R_e \rightarrow R_c$. The more the functions expressed by exponent e become similar to the universe of functions and meanings carried by the inflectional class, the less distinctive the exponent becomes. In other words, an exponent is more successful as a distinctive functional unit of the language when $R_c - R_e$ is large. If so, the amount of information $I'_{R_e|R_c}$ is large, and hence I_{w_e} in (20) is small, and as a consequence processing latencies are reduced. By contrast, an exponent for which $I'_{R_e|R_c}$ is small is dysfunctional, and therefore harder to process, leading to longer processing latencies.

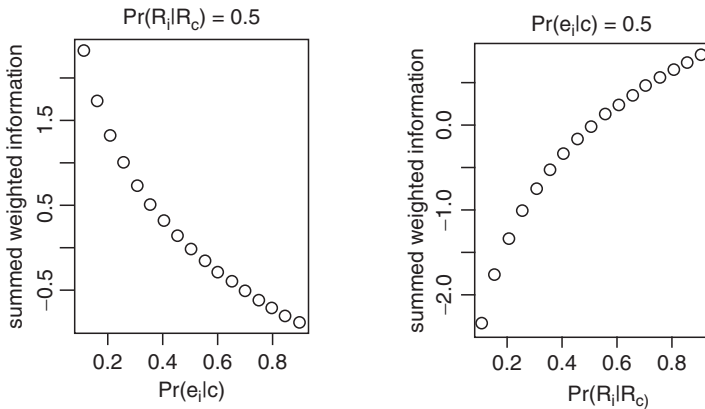


FIGURE 10.3 The left panel shows the partial effect of the information carried by the probability of the exponent given its class $I'_{e_i|c}$. The right panel shows the partial effect of the information carried by the proportion of the number of functions and meanings conditioned on the total number of functions and meanings for the class $I'_{R_e|R_c}$. Both partial effects are calibrated for the other effect evaluated at 0.5, and are calculated straightforwardly from (20).

10.4 The information structure of paradigms

10.4.1 Entropy

Thus far, we have considered the processing load of an inflected form given its paradigm, or an exponent, given its inflectional class. Moscoso del Prado Martín *et al.* (2004b) added a new dimension to the experimental study of morphological connectivity by considering the cost of the complexity of a paradigm as such, gauged by means of the entropy measure H . Figure 10.1 is helpful for discussing the difference between Kostić's approach and the one developed by Moscoso del Prado Martín and his colleagues. Ignoring the weighting for numbers of functions and meanings, Kostić's measure simplifies to $-\log_2(\text{Pr}_\pi(e))$, which reflects the number of steps from the root node to the leaf node of the exponent e in an optimal binary coding scheme (see the upper panel; for numbers of nodes that are integer powers of two, the $-\log_2(\text{Pr}_\pi(e))$ is exactly equal to the number of steps). However, this measure is insensitive to the size and configuration of the tree. To capture these aspects of the tree, we can make use of the entropy measure. The entropy, which is the same for each and every member of the paradigm, quantifies the expected number of steps from the root to a leaf node.

Moscoso del Prado Martín *et al.* (2004b) applied the entropy measure to paradigms in Dutch, but used a much broader definition of paradigms that extended the concept of the morphological family. Table 10.4 shows the words listed in CELEX that contain *neighbour* as a constituent. The left two columns list the morphological family as defined by Schreuder and Baayen (1997), the middle columns list the inflected variants that were found for two of the

TABLE 10.4 Morphological family and inflectional paradigms for *neighbour*.

morphological family		inflectional paradigms		merged paradigms	
word	F	word	F	word	F
<i>neighbour</i>	901	<i>neighbour</i>	343	<i>neighbour</i>	343
<i>neighbourhood</i>	407	<i>neighbours</i>	558	<i>neighbours</i>	558
<i>neighbouring</i>	203			<i>neighbourhood</i>	386
<i>neighbourliness</i>	3	<i>neighbourhood</i>	386	<i>neighbourhoods</i>	21
<i>neighbourly</i>	14	<i>neighbourhoods</i>	21	<i>neighbouring</i>	203
				<i>neighbourliness</i>	3
				<i>neighbourly</i>	14

members of the family, and the rightmost columns list the set that merges the family members with the inflected variants. Moscoso del Prado Martín and colleagues calculated the entropy over this merged set, and proposed this entropy as an enhanced measure for capturing the morphological family size effect. They pointed out that, when all family members are equiprobable, the entropy of the family reduces to the log of the number of family members. Since it is exactly this log-transformed count that emerged as predictive for processing latencies, the entropy of the family can be viewed as a principled way of weighting family members for their token frequency.

Moscoso del Prado Martín and colleagues combined this generalized entropy measure with the amount of information carried by a word (inflected or uninflected) as estimated from its relative frequency to obtain what they called the information residual:

$$I_R = I_w - H = \log N - \log_2 F_w - H. \quad (21)$$

This information residual performed well in a series of post-hoc analyses of processing of Dutch complex words.

By bringing several measures together in a single predictor, I_R , stem frequency and entropy receive exactly the same regression weight:

$$\begin{aligned} RT &\propto \beta_0 + \beta_1 I_R \\ &= \beta_0 + \beta_1 (I_w - H) \\ &\quad \beta_0 - \beta_1 \log_2 F_w - \beta_1 H. \end{aligned} \quad (22)$$

However, subsequent work (Baayen *et al.* 2006) suggests that frequency, the entropy calculated over the morphological family while excluding inflected variants, and the entropy of the paradigms of individual lexemes should be allowed to have different importance (i.e, different β weights). Their study examined a wide range of lexical predictors for simple English nouns and verbs, and observed independent effects of inflectional entropy (henceforth H_i) across both the visual lexical decision and word-naming tasks. An effect of derivational entropy (henceforth H_d) was present only in the visual lexical decision task. Here, it emerged with a U-shaped curve, indicating the presence of some inhibition for words with very information-rich families. In their study of the lexical processing of 8486 complex words in English, Baayen *et al.* (2008c) also observed an independent facilitatory effect of inflectional entropy, side by side with a facilitatory effect of the family size of the lexeme.

These results suggest that, when considered in terms of optimal binary coding schemes, inflected words and lexemes should not be brought together in one encompassing binary tree. Instead, lexemes form one tree, and each lexeme then comes with its own separate disjoint tree for its inflected variants.

Inflectional paradigms in languages such as Dutch and English are trivially simple compared to the paradigms one finds in morphologically rich languages. This raises the question to what extent entropy measures inform us about the processing complexity of more substantive paradigmatic structure. We address this issue for nominal paradigms in Serbian.

10.4.2 *Relative entropy*

When the inflectional entropy is computed for a given lexeme, it provides an estimate for the complexity of this lexeme's inflectional paradigm. This measure, however, does not take into account the complexity of the inflectional class, and the extent to which the probability distribution of a lexeme's paradigm diverges from the probability distribution of its inflectional class. We could consider bringing the entropy of the inflectional class into our model, but this class entropy would be the same for all lexemes in the class. Hence, it would not be much more informative than a plain name for that class (for example, Latin declension I, or Serbian declension III). Therefore, Milin *et al.* (2009) considered the simultaneous influence of paradigms and classes on the processing of inflected nouns in Serbian by means of relative entropy, *RE*.

Milin *et al.* (2009) investigated whether relative entropy is predictive for lexical processing in visual lexical decision using masculine and feminine nouns with the case endings *-om*, *-u*, and *-e*. A mixed-effects analysis with word frequency and stem frequency, bigram frequency, number of orthographic neighbors and entropy as covariates revealed an independent inhibitory effect of *RE*, as shown in the lower right panel of Figure 10.4. Comparison with the other significant partial effects in the model shows that the magnitude of the effect of *RE* is comparable to that of stem frequency and orthographic neighborhood size. However, the effect of the entropy did not reach significance ($p > 0.15$).

What this experiment shows is that it is neither the probability distribution of the inflected variants in a word's paradigm, nor the probability distribution in its inflectional class considered separately that are at issue, but rather the divergence between the two distributions. The greater this divergence, the longer the response latencies. A similar pattern was observed for the accuracy measure as well: the greater the divergence of the probability distribution of the paradigm from the probability distribution of the class, the more errors were made.

From the perspective of cognitive psychology, these results are interesting in that they provide further evidence for the importance of structured

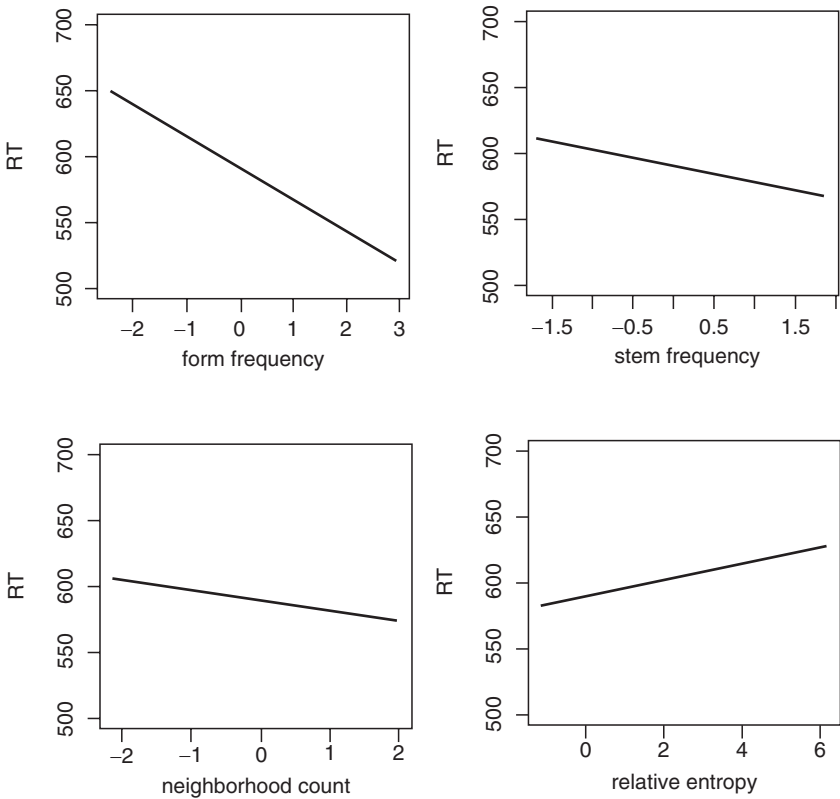


FIGURE 10.4 Partial effects of distributional predictors for the response latencies in visual lexical decision to Serbian nouns (Milin *et al.* 2008)

lexical connectivity. From the perspective of linguistic morphology, they support the theoretical concepts of paradigms and inflectional classes. Combined with the presence of a strong effect of the word frequency, an effect that is much stronger than the effect of the word's stem (compare the upper panels in Figure 10.4), these results provide strong support for word and paradigm morphology (Matthews 1974; J. P. Blevins 2003, 2006*b*) and for exemplar-based approaches to lexical processing in general (see, e.g., Baayen 2003).

10.5 Paradigmatic structure in derivation

In languages such as Dutch or English, morphological families consist predominantly of compounds. As a consequence, the family size effect (cf., Schreuder and Baayen 1997) is driven almost exclusively by lexical

TABLE 10.5 The number of monomorphemic base words that can attach the given number of affixes (prefixes or suffixes) when forming bi-morphemic derived words.

Number of affixes	Count of base words
1	3449
2	1391
3	516
4	202
5	105
6	31
7	13
8	11
9	2
10	3
11	2

connectivity between compounds. Little is known about the role of derived words. The problem here is that a given base word combines with only a handful of derivational affixes at best. Counts of the number of different prefixes and suffixes that English monomorphemic base words combine with, based on the English section of the CELEX lexical database (Baayen *et al.* 1995), illustrate that 60 percent of English monomorphemic base words combine with only one affix. Table 10.5 shows a steep decrease (a Zipfian distribution) in the number of derivational affixes that are attested for a given base word. The verbs *act* and *play* are exceptional in combining with 11 different affixes. The maximum family size in English, 187, observed for *man*, is an order of magnitude larger. With such small numbers of derived family members, it becomes very difficult to gauge the role of a strictly derivational family size count in lexical processing.

Derived words, however, enter into more systematic relations than most compounds, even when we take into account that the meaning of a compound is predictable from its constituents to a much greater extent than has traditionally been assumed (Gagné and Shoben 1997; Gagné 2001). For instance, derived adjectives with the prefix *un-* systematically express negation. Taking this fact into account, we asked ourselves whether such systematic relations between base words and their derivatives codetermine lexical processing. As a first step towards an answer, we introduce two simple concepts: the mini-paradigm and the mini-class. Here, the term mini-paradigm refers to pairs of base words and their derivatives. Thus, *kind* and *unkind* form a mini-paradigm, and so do *clear* and *clearly*. In the same line, the term

mini-class refers to the set of mini-paradigms sharing the same derivational affix. All pairs of base words and the corresponding *un-* derivatives constitute the mini-class of: *kind - unkind, true - untrue, pleasant - unpleasant*, etc. Mini-paradigms and mini-classes approximate inflectional paradigms and inflectional classes in the sense that the semantic relations within the pairs tend to be more consistent and transparent than in general morphological families or in families of derived words with different prefixes and suffixes.

In what follows, we therefore investigate whether the measures of entropy and relative entropy are significant predictors for lexical processing when applied to mini-paradigms and mini-classes.

10.5.1 *Materials*

We selected six suffixes and one prefix, for which we extracted all formations listed in the CELEX lexical database and for which latencies were also available in the English Lexicon Project (Balota et al. 2007) for both the derived word and its base. The resulting counts of formations are available in Table 10.6, cross-classified by whether the base word is simple or complex. For all words, we extracted from CELEX their frequency of occurrence, their length in letters, the number of synsets for the base as listed in WordNet (Miller 1990; Beckwith et al. 1991, and studied by Baayen et al. 2006), the family size of the base (calculated from the morphological parses in CELEX), and their frequency in the demographic subcorpus of conversational English in the British National Corpus (Burnard 1995). We included these variables in order to make sure that potential paradigmatic effects are not confounded with other lexical distributional properties. From the English Lexicon Project, we added the by-item mean naming latencies and the by-item mean lexical decision latencies.

TABLE 10.6 Affixes in the study based on latencies extracted from the English Lexicon Project, cross-classified by the complexity of their base words.

	simple base	complex base
<i>-able</i>	70	0
<i>-er</i> (comparative)	98	0
<i>-er</i> (deverbal)	240	24
<i>-ly</i> (adverbial)	21	355
<i>-ness</i> (complex base)	0	65
<i>-ness</i> (simple base)	152	0
<i>-est</i> (superlative)	95	0
<i>un-</i>	18	111

For each pair of base and derivative, we calculated its entropy and its relative entropy. For the derived words, the entropy of the mini-paradigm was calculated on the basis of the relative frequencies of the derivative and its base word (e.g., for *kind* and *unkind*, the relative frequencies are $72/(72 + 390)$ and $390/(72 + 390)$). For the base words, we distinguished between base words with only one derivative, and base words with two or more derivatives. For base words with a single derivative, the procedure for estimating the entropy was the same as for derived words. For base words with more than one derivative, the problem arises how to calculate entropies. Selection of a single derivative seems arbitrary. Taking all derivations linked with a given base word into account is possible, but then the mini-class distribution would contain the maximum number of eleven relative frequencies (see Table 10.5), most of which would be zero for almost all words. We therefore opted for taking only two relative frequencies into account when calculating the entropy: the frequency of the base itself, and the summed frequency of all its derivatives.

The probability distribution for a given mini-class was obtained by summing the frequencies of all base words in the class on the one hand, and all derivatives in the class on the other hand. The resulting frequencies were then transformed into relative frequencies. These relative frequencies then served as the Q distribution (also known as the reference distribution) for the calculation of the relative entropy.

In the following analyses, frequency measures, family size, number of synsets, and response latencies were log-transformed to eliminate the adverse effect of outliers on the model fit.

10.5.2 *Derived words*

We investigated the predictivity of the entropy and relative entropy measures for word naming and lexical decision latencies to the derived words. For that, we applied linear mixed-effects modeling (Baayen *et al.* 2008a; Bates 2005, 2006; Baayen 2008), with Task (lexical decision versus naming) as a fixed-effect factor, and with the set of relevant covariates including length, (written) base frequency, (written) word frequency, spoken word frequency, number of synsets in WordNet, morphological family size, entropy and relative entropy. Word and affix were considered as random effects.

For the covariates, we investigated whether nonlinearity was present. This turned out to be the case only for word length. We also observed interactions of Task with word frequency and spoken word frequency, with length (only the quadratic term), and with entropy and relative entropy. Finally, we considered whether by-word or by-affix random slopes were required. It

turned out that by-affix random slopes were necessary only for the two entropy measures.

Inspection of the coefficients for the entropy measures in the resulting model revealed that entropy and relative entropy had positive coefficients of similar magnitude ($H : 0.034, \hat{\sigma} = 0.025$; $RE : 0.058, \hat{\sigma} = 0.016$), with small differences across the two tasks. In word naming, the effect of entropy was slightly larger, while the effect of relative entropy was fractionally smaller (H in naming: $0.034 + 0.041$; RE in naming: $0.058 - 0.014$).

These observations invite a simplification of the regression model. Let β_0 denote the coefficient for the intercept, and let β_1 and β_2 denote the coefficients for entropy and relative entropy respectively. Given that β_1 and β_2 are very similar, we can proceed as follows:

$$\begin{aligned} \beta_0 + \beta_1 H + \beta_2 RE &\approx \beta_0 + \beta_1 H + \beta_1 RE \\ &= \beta_0 + \beta_1 (H + RE). \end{aligned} \tag{23}$$

Interestingly, the sum of entropy and relative entropy is equal to another information- theoretic measure, the *cross entropy* (CE) (Manning and Schütze 1999; Cover and Thomas 1991). Applied to the present data, we have

$$\begin{aligned} CE &= H + RE = \\ &= - \sum_L \Pr_\pi(w_L) \log_2 (\Pr_\pi(w_L)) + RE \\ &= - \sum_L \Pr_\pi(w_L) \log_2 (\Pr_\pi(w_L)) + \sum_L \Pr_\pi(w_L) \log_2 \frac{\Pr_\pi(w_L)}{\Pr_\pi(c_L)} \\ &= - \sum_L \Pr_\pi(w_L) \log_2 (\Pr_\pi(c_L)). \end{aligned} \tag{24}$$

In (24), L indexes the base and derived lexemes for mini-paradigms, and the sets of base words and derived words for the mini-class. Thus, $\Pr_\pi(w_L)$ denotes the probability of a base or derived lexeme in its mini-paradigm, and $\Pr_\pi(c_L)$ denotes the corresponding probability in the mini-class. Technically, the cross entropy between the probability distribution of the mini-paradigm and the probability distribution of the mini-class measures the average number of bits needed to identify a form from the set of possible forms in the mini-paradigm, if a coding scheme is used based on the reference probability distribution $\Pr_{\pi c_e}$ of the mini-class, rather than the “true” distribution $\Pr_{\pi w_e}$ of the mini-paradigm. More informally, we can interpret the cross entropy as gauging the average amount of information in the

TABLE 10.7 Partial effects of the predictors for the visual lexical decision and naming latencies to derived words. The reference level for Task is lexical decision. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

	Estimate	Lower	Upper	P
Intercept	6.6679	6.5830	6.7607	0.0001
Task=naming	-0.1419	-0.2158	-0.0688	0.0001
length (linear)	0.0056	-0.0109	0.0228	0.5162
length (quadratic)	0.0012	0.0004	0.0020	0.0034
written frequency	-0.0382	-0.0428	-0.0333	0.0001
spoken frequency	-0.0183	-0.0245	-0.0117	0.0001
synset count	-0.0277	-0.0339	-0.0212	0.0001
cross entropy	0.0565	0.0164	0.0937	0.0076
Task=naming: written frequency	0.0067	0.0022	0.0112	0.0036
Task=naming: length (linear)	0.0132	-0.0025	0.0283	0.0914
Task=naming: length (quadratic)	-0.0011	-0.0019	-0.0003	0.0026
Task=naming: spoken frequency	0.0124	0.0062	0.0186	0.0001

mini-paradigm, corrected for the departure from the prior reference distribution of the corresponding mini-class.

We therefore replaced entropy H and relative entropy RE as predictors in our regression model by a single predictor, the cross entropy CE , and refitted the model to the data. After removal of outliers and refitting, we obtained the model summarized in Table 10.7 and visualized in Figure 10.5. The standard deviation of the by-word random intercepts was 0.0637, the standard deviation for the by-affix random intercepts was 0.0399, the standard deviation for the by-affix random slopes for cross entropy was 0.0277, and the standard deviation for the residual error was 0.0663. All random slopes and random intercepts were supported by likelihood ratio tests (all p-values < 0.0001).

With respect to the control variables, we note that word length was a strongly nonlinear (positively accelerated) predictor for especially lexical decision, with longer lengths eliciting elongated response latencies. The word frequency effect was similar for both tasks, albeit slightly stronger for lexical decision. Similarly, the spoken word frequency added facilitation specifically for lexical decision. The effect of number of synonyms, as gauged with the help of the synset count, was facilitatory and the same across the two tasks. The effect of cross entropy was inhibitory, and also did not differ across tasks. Its effect size (roughly 100 ms) exceeds that of the spoken frequency effect and that of the number of meanings. Interestingly, the model with cross entropy as predictor provides an equally tight fit to the data as the model with entropy and relative entropy as predictors, even though the latter model had

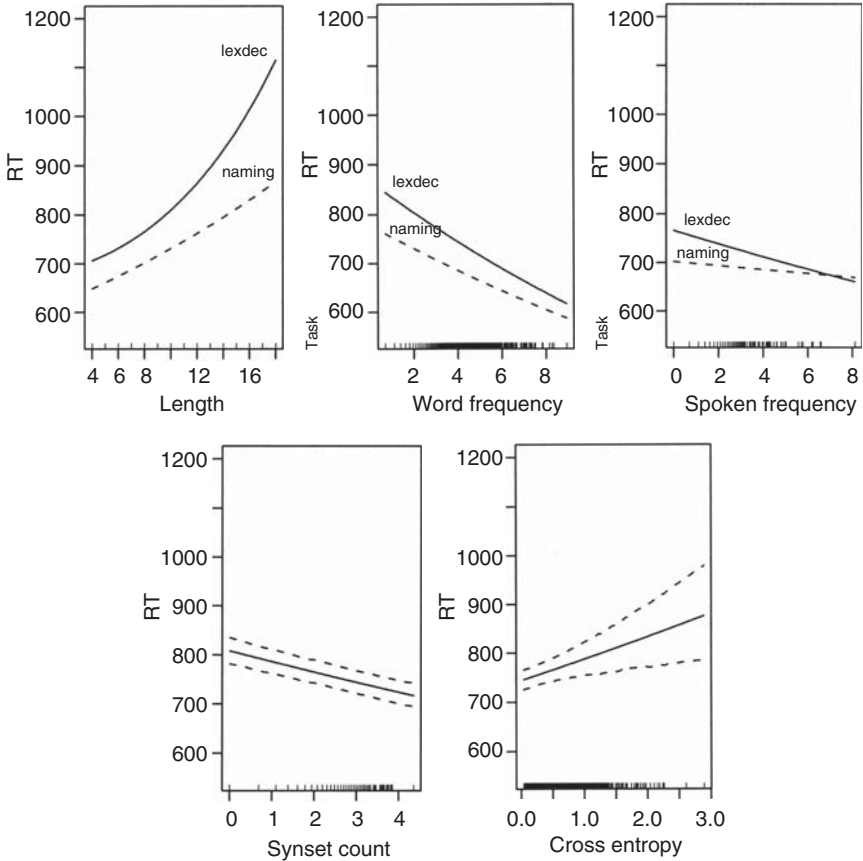


FIGURE 10.5 Partial effects of the predictors for word naming and visual lexical decision latencies for derived words. The lower panels are calibrated for visual lexical decision, and come with 95% highest posterior density confidence intervals.

two additional parameters (a beta coefficient for a second entropy measure, and a random-effects standard deviation for by-item slopes for the second entropy measure): the log likelihood of the simpler model with cross entropy was 2364, while for the more complex model with entropy and relative entropy it was 2362.³ From this, we conclude that the relevant entropy measure for understanding the role of paradigmatic complexity during lexical processing of derived words is the cross-entropy measure.

³ A greater log likelihood implies a better fit (for technical details consult Crawley 2002).

TABLE 10.8 Estimated slopes for derived words for the different mini-classes, positioned in decreasing order.

	slope
<i>-est</i> (superlative)	0.097
<i>-ly</i> (adverbial)	0.090
<i>-ness</i> (complex base)	0.086
<i>-able</i>	0.068
<i>-er</i> (comparative)	0.054
<i>-er</i> (deverbal)	0.031
<i>un-</i>	0.021
<i>-ness</i> (simple base)	0.004

The synset measure in our data estimates the number of meanings that a base word has (e.g., *bank* as a part of the river and a financial institution). Generally, the meaning of a derivative builds on only one of the meanings of its base word (e.g., *embank*). The lower the number of synsets, the tighter we may expect the relationship between the base and its derivatives to be. The synset measure does not interact with cross entropy, nor does it substantially affect the estimate of its slope. To further rule out potential semantic confounds, we also considered a semantic measure that specifically gauges the semantic similarity between a given derived word and its base. The measure that we used in the LSA score for the distance between the derived word and its base in co-occurrence space (Landauer and Dumais 1997), using the software available at <http://lsa.colorado.edu>. For the subset of our mini-paradigms, the LSA scores elicited a significant facilitatory effect on lexical decision latencies ($\beta = -0.1196$, $p = 0.0001$). As for the synset measure, there was no significant effect for word naming. Crucially, the measure of cross entropy retained significance also when the pairwise semantic similarity between base and derived word in mini-paradigms was taken into account.

The presence of random slopes for cross entropy in this model indicates that the effect of cross entropy varied with mini-class. Table 10.8 lists the individual slopes for the different mini-classes that we considered. Slopes range from 0.097 for superlative *-est* to 0.004 for *-ness* formations derived from simple base words.

10.5.3 Base words

Because complex base words (e.g., *surprising*) come with predictors such as the frequency of the stem (*surprise*) that do not apply to the simple base words, we analyzed the simple and complex base words separately. We

TABLE 10.9 Partial effects of the predictors for word naming and visual lexical decision latencies for complex base words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

	Estimate	Lower	Upper	P
Intercept	6.6006	6.5428	6.6596	0.0001
experiment=naming	-0.0397	-0.0750	-0.0031	0.0326
length	0.0357	0.0325	0.0387	0.0001
word frequency	-0.0305	-0.0363	-0.0250	0.0001
spoken frequency	-0.0143	-0.0195	-0.0090	0.0001
base frequency	-0.0061	-0.0086	-0.0035	0.0001
synset count	-0.0230	-0.0311	-0.0147	0.0001
cross entropy	-0.1038	-0.1605	-0.0483	0.0002
Experiment=naming: length	-0.0082	-0.0115	-0.0052	0.0001
Experiment=naming: word frequency	0.0100	0.0057	0.0141	0.0001

proceeded in the same way as for the derived words. We fitted a mixed-effects model to the data, observed that again the coefficients for entropy and relative entropy were very similar and statistically indistinguishable in magnitude and had the same sign, replaced the two measures by the cross-entropy measure, refitted the model, and removed overly influential outliers.

The coefficients of a mixed-effects model fitted to the lexical decision and naming latencies to the complex base words are listed in Table 10.9. The corresponding partial effects are graphed in Figure 10.6.

As for the preceding datasets, we find effects of word length (longer words elicit longer latencies, upper left panel) and word frequency (more frequent words elicit shorter latencies, uppercenter panel). Adding frequency of use in spoken English as a predictor again contributes significantly to the model over and above the written frequency measures (upper right panel). The frequency of the base word (lower left panel of Figure 10.6) also emerged as a significant predictor, but with a slope that is substantially shallower than that of the word frequency effect. The synset count of the embedded base word is predictive as well. It is facilitatory, just as observed for the derived words (lower center panel). Finally, the lower right panel shows that there is a small effect of cross entropy. But while for the derived words the effect of cross entropy was inhibitory, it is facilitatory for the base words.

Before discussing this unexpected change in sign, we first inquire whether facilitation for cross entropy also characterizes the set of simple base words. Table 10.10 lists the partial effects of the predictors that were retained after

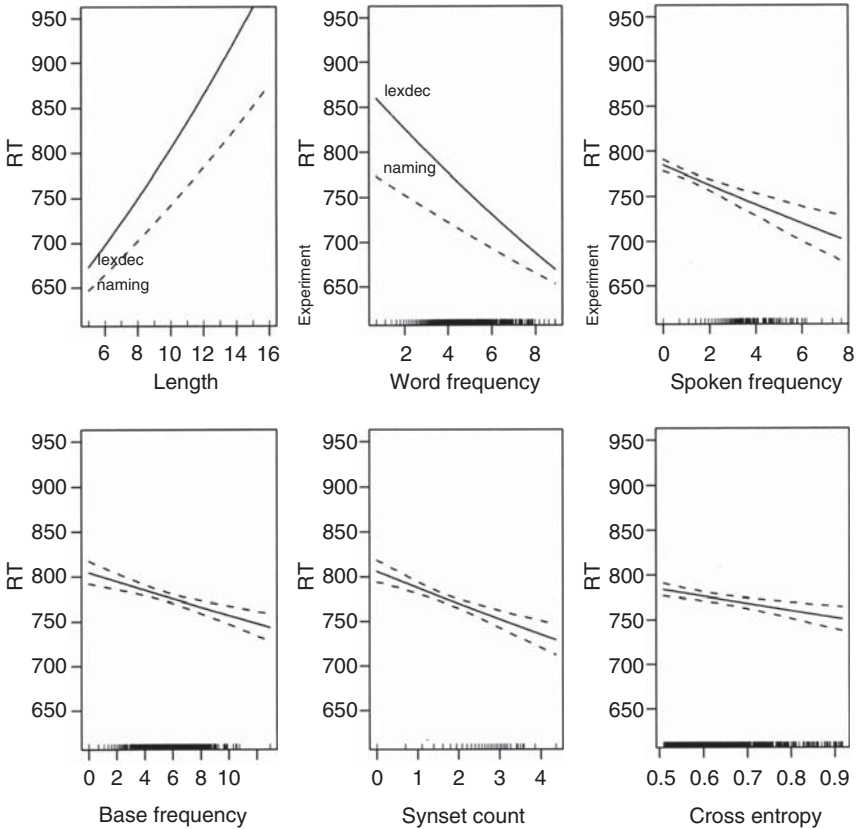


FIGURE 10.6 Partial effects of the predictors for word naming and visual lexical decision latencies for complex base words. Markov chain Monte Carlo based 95% confidence intervals are shown for those predictors that do not enter into interactions.

stepwise variable elimination. Figure 10.7 visualizes these partial effects. The upper left panel shows the effect of orthographic length, which shows a clear minimum near the median length (five letters) for visual lexical decision but not for word naming. For the latter task, the shorter the word, the easier it is to articulate. For the former task, five-letter words emerge as most easily read. The upper right panel shows that, as for the derived words, spoken frequency allows greater facilitation for visual lexical decision than for word naming.

The lower left panel presents the expected facilitatory effect of the synset count, and illustrates that words with more meanings elicit shorter latencies, for both word naming and lexical decision. Surprisingly, the lower central panel shows that the partial effect of family size is inhibitory, instead of facilitatory, as reported for previous experiments. We return to this finding

TABLE 10.10 Partial effects of the predictors for word naming and visual lexical decision latencies for simple base words. Lower, Upper: 95% highest posterior density interval; P: Markov chain Monte Carlo p-value.

	Estimate	Lower	Upper	P
Intercept	6.8433	6.7756	6.9097	0.0001
experiment=naming	-0.2520	-0.3213	-0.1885	0.0001
length (linear)	-0.0613	-0.0797	-0.0430	0.0001
length (quadratic)	0.0067	0.0052	0.0080	0.0001
spoken frequency	-0.0251	-0.0286	-0.0216	0.0001
family size	0.0107	0.0021	0.0193	0.0158
word frequency	-0.0090	-0.0125	-0.0054	0.0001
cross entropy	-0.1316	-0.1823	-0.0869	0.0001
synset count	-0.0235	-0.0321	-0.0154	0.0001
Experiment=naming: length (linear)	0.0507	0.0305	0.0722	0.0001
Experiment=naming: length (quadratic)	-0.0034	-0.0050	-0.0018	0.0002
Experiment=naming: spoken frequency	0.0173	0.0141	0.0202	0.0001

below. The partial effect of cross entropy is presented in the lower right panel of Figure 10.7. As for the complex base words, the effect of cross entropy for simple base words is again facilitatory.

The analyses of the two sets of base words leave us with two questions. First, how should we understand the change in sign of the cross-entropy effect between derived words and base words? Second, why do we have inhibition from the morphological family size for simple base words, and no effect of family size for complex base words?

With respect to the first question, we note that there is bottom-up support for only the base word, and no such support for their derivatives. By contrast, in the case of the derived words, there is bottom-up support for the derived word itself, its base word, and its affix. In sum, for derived words, three of the four elements in a proportional analogy such as

$$\underbrace{\text{great} : \text{greatest}}_{\text{mini-paradigm}} = \underbrace{\text{A} : \text{-est}}_{\text{mini-class}} \quad (25)$$

are actually present in the signal. For derived words, we can therefore understand the effect of cross entropy as reflecting the cost of resolving the proportional analogy between mini-paradigm and mini-class. More specifically, the cross entropy reflects the average complexity of identifying the derived word in its mini-paradigm on the basis of the generalized probability distribution of the

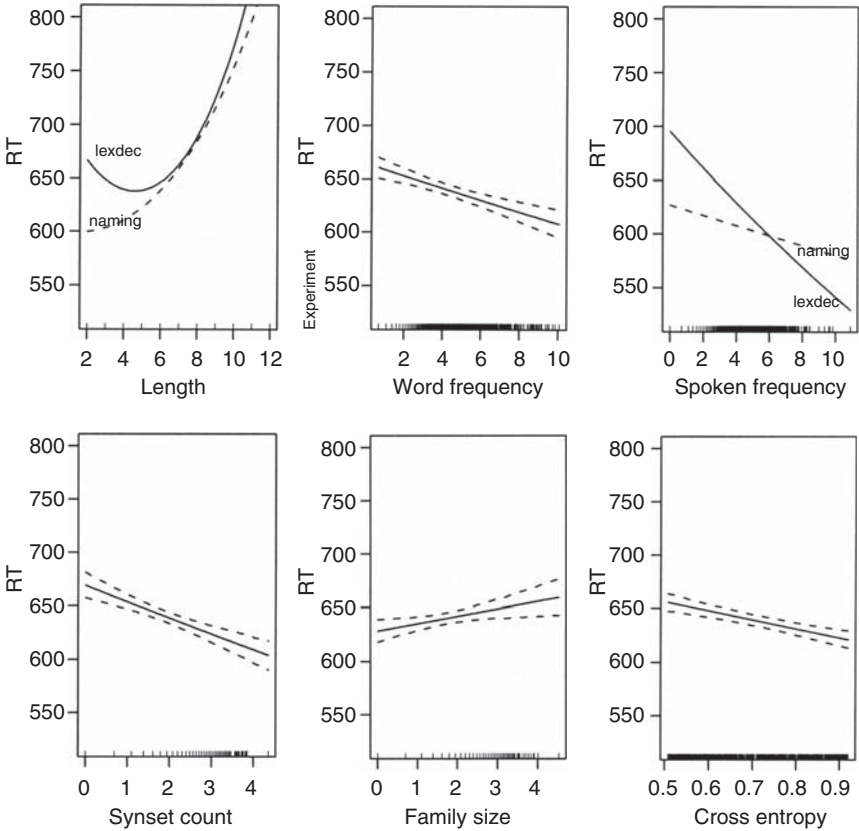


FIGURE 10.7 Partial effects of the predictors for word naming and visual lexical decision latencies for simple base words. Markov chain Monte Carlo based 95% confidence intervals are shown for those predictors that do not enter into interactions.

mini-class. Thus, the cross entropy can be understood as reflecting the cost of resolving the ambiguity in the visual input with the help of generalized knowledge in long-term memory about the corresponding mini-class. From this perspective, the inhibitory effect of cross entropy for derived words makes perfect sense: The higher the cross entropy, the more information has to be retrieved from memory to resolve the proportional analogy.

Let us now consider the facilitatory effect of cross entropy for simple base words. For simple base words, the visual input is unambiguous, with bottom-up support only for the word itself. There is no cost of a call on proportional analogy to resolve morphological ambiguity. In the absence of a

TABLE 10.11 Pairwise correlations between key predictors and lexical decision (lexdec) and naming latencies for the set of simple base words.

	Frequency	Family size	Synset count	Cross entropy	RT lexdec	RT naming
frequency	1.000	0.320	0.345	-0.527	-0.379	-0.266
family size	0.320	1.000	0.643	0.245	-0.473	-0.392
synset count	0.345	0.643	1.000	0.092	-0.552	-0.434
cross entropy	-0.527	0.245	0.092	1.000	-0.085	-0.101
RT lexical decision	-0.379	-0.473	-0.552	-0.085	1.000	0.648
RT naming	-0.266	-0.392	-0.434	-0.101	0.648	1.000

morphological parsing problem, the cross-entropy effect apparently reverses and emerges as a measure of the amount of support the base receives from related derived words co-activated by the base. Crucially, it is not simply the count of related derived words (we checked that this count is not predictive for the present data) but rather the analogical support for the base given its derivative (defined in the mini-paradigm) and the general likelihood of a base word having derivatives (defined in the miniclass).

The second question to be considered is why we observe inhibition from the morphological family size for simple base words, and no effect of family size for complex base words. The unexpected inhibitory effect of family size is probably due to what is known in the statistical literature as suppression (see, e.g., Friedman and Wall 2005): When predictor variables are correlated, and both are correlated with the dependent variable, then, depending on the strength of the former correlation, the beta coefficient of one of the predictors can become nonsignificant or even change sign. Table 10.11 presents the correlation matrix for key predictors, and reveals a large positive coefficient for the correlation of family size and the synset count, and the expected negative correlations for family size and response latencies in lexical decision and naming. This by itself is a warning that suppression might be at issue here.

We therefore inspected whether family size was significant in a model for the simple base words, excluding the synset count as predictor. It was not ($p > 0.8$). When cross entropy was also removed as predictor, the family size measure emerged as significant ($p < 0.01$), now with a negative slope, as expected given previous studies. For the complex base words, excluding only the synset measure was sufficient to allow a facilitatory effect of family size to emerge. What this suggests is that the family size effect, which has always been understood as a semantic effect (see, e.g., Schreuder and Baayen 1997; Moscoso del Prado Martín *et al.* 2004a), is a composite effect that bundles

effects of semantic similarity and effects of paradigmatic structure. Effects of similarity would then be better captured by means of the synset count, and effects of derivational paradigmatic structure would then be better captured by means of the cross-entropy measure.

The question that arises at this point is whether the semantic aspect of the family size effect has any specific morphological component. To answer this question, we first partitioned the synset count into two disjunct counts, a count for morphologically related synsets, and a count for morphologically unrelated synsets. A morphologically related synset is a synset in which at least one of the synset members is morphologically related to the target word (not counting the target word itself). A morphologically related synset, therefore, is a family size count that only includes semantically highly related family members.

In the model for the simple base words, we then replaced the family size measure and the synset count by the counts of morphologically related and unrelated synset counts. A mixed-effects analysis revealed that, for visual lexical decision, both counts were significant predictors with very similar coefficients (-0.018 and -0.015 respectively). For the naming latencies, however, only the synset count of morphologically unrelated synsets was significant. This interaction ($p = 0.0049$) shows that in a task such as word naming, which does not require deep semantic processing, semantic ambiguity that arises through morphological connectivity does not play a role. By contrast, the lexical decision task, which invites deeper semantic processing, allows the effect of morphologically related words that are also very similar in meaning to become visible. We therefore conclude that morphologically related words that are also semantically very similar have a special status compared to semantically similar but morphologically unrelated words (see also Moscoso del Prado Martín *et al.* 2004a).

10.6 Concluding remarks

In the preceding sections we reviewed and presented a range of studies addressing specific aspects of the complexities of paradigmatic structure in lexical processing. In order to obtain a model for the full complexity for an inflected variant w_e , we combine equations (10), (14), and (15) and add the effects of the entropy and relative entropy measures, leading to the following equation:

$$\begin{aligned}
I \propto & \beta_0 + \beta_1 \log_2 \Pr_N(w_e) + \beta_2 \log_2 \Pr_N(w) + \\
& + \beta_3 \log_2 \left(\frac{\Pr_\pi(e)/R_e}{\sum_e \Pr_\pi(e)/R_e} \right) + \\
& + \beta_4 \log_2 \left(\frac{\Pr_\pi(w_e)/R_e}{\sum_e \Pr_\pi(w_e)/R_e} \right) + \\
& + \beta_5 H_d + \\
& + \beta_6 H_i + \beta_7 RE.
\end{aligned} \tag{26}$$

Large regression studies are called for to bring all these variables into play simultaneously. However, even though (26) is far from simple, it is only a first step towards quantifying the complexities of inflectional processing. We mention here only a few of the issues that should be considered for a more comprehensive model.

First, Kostić *et al.* (2003) calculated the number of functions and meanings R_e of exponent e conditionally on a lexeme's inflectional class. For instance, the number of functions and meanings listed for the exponent a for masculine nouns in Table 2, 109, is the sum of the numbers of functions and meanings for masculine genitive and the masculine accusative singular. This provides a lower bound for the actual ambiguity of the exponent, as the same exponent is found for nominative singulars and genitive plurals for regular feminine nouns. The justification for conditioning on inflectional class is that the stem to which an exponent attaches arguably provides information about its inflectional class. This reduces the uncertainty about the functions and meanings of an exponent to the uncertainty in its own class. Nevertheless, it seems likely that an exponent that is unique to one inflectional class (e.g., Serbian *ama* for regular feminine nouns) is easier to process than an exponent that occurs across all inflectional classes (e.g., *a*, *u*), especially when experimental items are not blocked by inflectional class. (Further complications that should be considered are the consequences of, for instance, masculine nouns (e.g., *sudija* ('judge'), *sluga* ('servant')) taking the same inflectional exponents as regular feminine nouns do, and of animate masculine nouns being associated with a pattern of exponents that differs from that associated with inanimate masculine nouns.)

Second, the standard organization of exponents by number and case has not played a role in the studies that we discussed. Thus far, preliminary analyses of the experimental data available to us have not revealed an independent predictive role for case, over and above the attested role of ambiguity with respect to numbers of functions and meanings. This is certainly an issue that requires further empirical investigation, as organization by case provides

insight into the way that functions and meanings are bundled across inflectional classes.

Third, we have not considered generalizations across, for instance, irregular and regular feminine nouns in Serbian, along the lines of Clahsen *et al.* (2001). The extent to which inflected forms inherit higher-order generalizations about their phonological form provides further constraints on lexical processing.

Fourth, the size of inflectional paradigms has not been investigated systematically. Although the nominal inflectional classes of Serbian are an enormous step forward compared to the nominal paradigms of English or Dutch, the complexities of verbal paradigms can be much larger. From an information-theoretic perspective, the entropy of the complex verbal paradigms of Serbian must be much larger than the entropy of nominal paradigms, and one would expect this difference to be reflected in elongated processing latencies for inflected verbs. The study by Traficante and Burani (2003) provides evidence supporting this prediction. They observed that inflected verbs in Italian elicited longer processing latencies than inflected adjectives.

Nevertheless, it should be noted that the question of what constitutes a verbal paradigm is still open. In one, traditional, sense each verb may have not one, but several paradigms defined over various tenses and aspects. In the other sense, verbs have one exhaustive paradigm that encompasses all verbal inflected variants. Baayen *et al.* (2008c) have addressed a similar question for the paradigms of English nouns and they concluded that lexemes and their inflected variants should not be considered together as a single paradigm. In a similar way, we can tackle the question of verbal paradigmatic organization in the mental lexicon – using information theory and large-scale regression modeling. Two alternatives can be tested empirically and the result should be straightforwardly in favor of either (a) a single entropy measure calculated over all verbal inflected variants or (b) entropies within each tense and aspect, and one computed over all tenses and aspects.

Fifth, all results reported here are based on visual comprehension tasks (lexical decision, word naming). Some of the present results are bound to change as this line of research is extended to other tasks and across modalities. For instance, the effect of inflectional entropy reported by Baayen *et al.* (2006) for visual lexical decision and word naming was facilitatory in nature. However, in a production study by Bien (2007), inflectional entropy was inhibitory (see also Baayen *et al.* 2008b). In lexical decision, a complex paradigm is an index of higher lexicality, and may therefore elicit shorter response latencies. In production, however, the paradigm has to be accessed, and a specific word form has to be extracted from the paradigm. This may explain why, in production, a greater

paradigm complexity appears to go hand in hand with increasing processing costs. Generally, it will be important to establish paradigmatic effects for lexical processing in natural discourse using tasks that do not, or only minimally, impose their own constraints on processing.

Sixth, it will be equally important to obtain distributional lexical measures that are more sensitive to contextual variation than the abstract frequency counts and theoretical concepts of functions and meanings that have been used thus far. Interestingly, Moscoso del Prado Martín *et al.* (2008) and Filipović Đurđević (2007) report excellent predictivity for lexical processing of more complex information-theoretic measures of morphological and semantic connectivity derived bottom-up from a corpus of Serbian.

It is clear that the information-theoretic measures that we have proposed and illustrated in this chapter capture only part of the multidimensional complexity of lexical processing. Hence, each measure can be understood as a plane cross-cutting this multidimensional space. In spite of these limitations, the extent to which the present information-theoretic approach converges with *wpm* is striking. Across our experimental datasets we find evidence for exemplars, irrespective of whether the language under investigation is Dutch, English, or Serbian. At the same time, we observe the predictivity of entropy measures, which generalize across probability distributions tied to subsets of these exemplars, and evaluate the complexity of paradigms and the divergence between different levels of morphological organization. However, all the results discussed here pertain to the processing of familiar words. In order to properly gauge the processing complexity of new inflected and derived words, it will be necessary to combine *wpm* and the present information-theoretic approach with computational models of language processing.

Such an integration is especially challenging because across computational models of linguistic generalization, whether abstractionist and implementing greedy learning (Albright and Hayes 2003), or memory-based and implementing lazy learning (Daelemans and Van den Bosch 2005; Keuleers *et al.* 2007; Keuleers 2008), a common finding is that it is type frequencies and not token frequencies on which generalization is based. In fact, type-based generalization has been found to be reflected in processing measures as well (see, e.g., Ernestus and Baayen 2004; Krott *et al.* 2004). Typically, current computational models (cf. Albright this volume) make use of much more sophisticated analogies than the traditional four-part analogy that we have referred to as a possible explanation for the effect of cross entropy.

To resolve this paradox, we note, first of all, that our hypothesis is not a hypothesis about the choice of a linguistic form, but rather a measure of the cost of selecting a given complex word from its mini-paradigm given its mini-class.

Furthermore, note that for most of the derivational suffixes we have considered, there are no rival suffixes comparable to the rivaling options that characterize the past tense in English (Albright and Hayes 2003), or plural selection in Dutch (Keuleers *et al.* 2007). There is only one way in English to express the comparative, the superlative, or adverbs through suffixation. Hence, the probability of the selection of *-er*, *-est*, or *-ly* is equal to one. For this “degenerate” case, four-part analogy provides a reasonable model. In fact, we think it is precisely this uniformity in the analogical support for a given suffix that allows us to see the effect of cross entropy. Because there are no competing sets of exemplars supporting different outcomes, there are no overriding type frequency effects. As a consequence, the more subtle relevance of the token counts becomes visible only for the basic, type-uniform four-part analogy. The real challenge for future research, therefore, is to clarify whether subtle effects of token frequencies also codetermine the fine details of lexical processing when more complex, type-frequency-driven analogies come into play.

References

- Ackerman, Farrell, Blevins, James P., and Malouf, Robert (2008). Inflectional morphology as a complex adaptive system. Paper presented at the First Workshop on Complex Systems and Language, University of Arizona.
- Akhtar, Nameera (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language*, 26: 339–56.
- Albright, Adam (2002a). *The Identification of Bases in Morphological Paradigms*. Ph.D. thesis, UCLA.
- (2002b). Islands of reliability for regular morphology: Evidence from Italian. *Language*, 78: 684–709.
- (2008). Explaining universal tendencies and language particulars in analogical change. See Good (2008), 144–81.
- Andrade, Argelia Edith, and Hayes, Bruce (2001). Segmental environments of Spanish diphthongization. In *UCLA Working Papers in Linguistics, Number 7: Papers in Phonology 5*, A. Albright and T. Cho (eds.). UCLA, Los Angeles, 117–51.
- and Hayes, Bruce (2002). Modeling English past tense intuitions with minimal generalization. In *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, M. Maxwell (ed.). ACL, 58–69.
- — (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, 90: 119–61.
- — (2006). Modeling productivity with the Gradual Learning Algorithm: The problem of accidentally exceptionless generalizations. In *Gradiance in Grammar: Generative Perspectives*, F. Gisbert, F. Caroline, V. Ralf, and S. Matthias (eds.). Oxford: Oxford University Press, 185–204.
- Alegre, Maria and Gordon, Peter (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language*, 40: 41–61.
- Alverson, Hoyt (1994). *Semantics and Experience: Universal Metaphors of Time in English, Mandarin, Hindi and Sesotho*. Baltimore: Johns Hopkins University Press.
- Anderson, Judi Lynn, Martínez, Isaac H., and Pace, Wanda J. (1990). Comaltepec Chinantec tone. See Merrifield and Rensch (1990), 3–20.
- Anderson, Stephen R. (1992). *A-Morphous Morphology*. Cambridge: Cambridge University Press.
- (2004). Morphological universals and diachrony. In *Yearbook of Morphology 2004*, G. Booij and J. van Marle (eds.). Dordrecht: Springer, 1–17.
- Anttila, Raimo (1977). *Analogy*. The Hague: Mouton.
- and Brewer, Warren A. (1977). *Analogy: A Basic Bibliography*. Amsterdam: John Benjamins.
- Aronoff, Mark (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, MA: MIT Press.

- Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. In *Yearbook of Morphology 1992*, G. Booij and J. van Marle (eds.). Dordrecht: Kluwer, 181–208.
- (2003). Probabilistic approaches to morphology. In *Probabilistic Linguistics*, R. Bod, J. Hay, and S. Jannedy (eds.). Cambridge: Cambridge University Press, 229–87.
- (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.
- Davidson, Doug J., and Bates, Douglas M. (2008a). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*.
- Feldman, Laurie, and Schreuder, Robert (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 53: 496–512.
- Levelt, Willem M. J., Schreuder, Robert, and Ernestus, Mirjam (2008b). Paradigmatic structure in speech production. *Chicago Linguistics Society*, 43, to appear.
- and Lieber, Rochelle (1991). Productivity and English derivation: A corpus-based study. *Linguistics*, 29: 801–43.
- — and Schreuder, Robert (1997). The morphological complexity of simple nouns. *Linguistics*, 35: 861–77.
- McQueen, James M., Dijkstra, Ton, and Schreuder, Robert (2003a). Dutch inflectional morphology in spoken- and written-word recognition. In *Morphological Structure in Language Processing*, R. H. Baayen and R. Schreuder (eds.). Berlin: Mouton de Gruyter.
- — — (2003b). Frequency effects in regular inflectional morphology: Revisiting Dutch plurals. In *Morphological Structure in Language Processing*, R. H. Baayen and R. Schreuder (eds.). Berlin: Mouton de Gruyter, 355–90.
- and Moscoso del Prado Martín, Fermín (2005). Semantic density and past tense formation in three Germanic languages. *Language*, 81: 666–98.
- Piepenbrock, Richard, and Gullikers, Léon (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia: University of Pennsylvania.
- Wurm, Lee H., and Aycok, Joanna (2008c). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *The Mental Lexicon*, 2: 419–63.
- Bailey, Todd M. and Hahn, Ulrike (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44: 568–91.
- Balota, David A., Yap, Melvin J., Cortese, Michael J., Hutchison, Keith I., Kessler, Brett, Loftis, Bjorn, Neely, James H., Nelson, Douglas L., Simpson, Greg, B., and Treiman, Rebecca (2007). The English lexicon project. *Behavior Research Methods*, 39: 445–59.
- Barnes, Jonathan and Kavitskaya, Darya (2002). Phonetic analogy and schwa deletion in French. In *Proceedings of the 27th Berkeley Linguistic Society*, 39–50.
- Bartens, Hans-Hermann (1989). *Lehrbuch der saamischen (lappischen) Sprache*. Hamburg: Helmut Buske Verlag.
- Bates, Douglas M. (2005). Fitting linear mixed models in R. *R News*, 5: 27–30.
- (2006). Linear mixed model implementation in lme4. Ms., Department of Statistics, University of Wisconsin, Madison.

- Bauer, Laurie (1998). When is a sequence of two nouns a compound in English? *English Language & Linguistics*, 2: 65–86.
- Beard, Robert (1995). *Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation*. Albany, NY: SUNY Press.
- Becker, Judith A. (1994). ‘sneak-shoes’, ‘swordsmen’ and ‘nose-beards’: A case study of lexical innovation. *First Language*, 14: 195–211.
- Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, and Miller, George A. (1991). WordNet: A lexical database organized on psycholinguistic principles. In *Lexical Acquisition. Exploiting On-Line Resources to Build a Lexicon*, U. Zernik (ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, 211–32.
- Bennett, Charles H. (1999). Quantum Information Theory. In *Feynman and Computation: Exploring the Limits of Computers*, Anthony J. G. Hey (ed.). Reading, MA: Perseus Books.
- Bergen, Benjamin K. (2004). The psychological reality of phonaestemes. *Language*, 80: 290–311.
- Berko, Jean (1958). The child’s learning of English morphology. *Word*, 14: 150–77.
- Berman, Ruth A. and Clark, Eve V. (1989). Learning to use compounds for contrast: data from Hebrew. *First Language*, 9: 247–70.
- Bertram, Raymond, Schreuder, Robert, and Baayen, R. Harald (2000). The balance of storage and computation in morphological processing. The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology-Learning Memory and Cognition*, 26: 489–511.
- Bien, Heidrun (2007). *On the Production of Morphologically Complex Words with Special Attention to Effects of Frequency*. Nijmegen: Max Planck Institute for Psycholinguistics.
- Blevins, James P. (2003). Stems and paradigms. *Language*, 79: 737–67.
- (2005). Word-based declensions in Estonian. In *Yearbook of Morphology 2005*, G. Booij and J. van Marle (eds.). Dordrecht: Springer, 1–25.
- (2006a). English inflection and derivation. In *Handbook of English Linguistics*, B. Aarts and A. M. S. McMahon (eds.). Oxford: Blackwell, 507–36.
- (2006b). Word-based morphology. *Journal of Linguistics*, 42: 531–73.
- (2007). Conjugation classes in Estonian. *Linguistica Uralica*, 43: 250–67.
- (2008). The post-transformational enterprise. *Journal of Linguistics*, 44: 723–42.
- Blevins, Juliette (2004). *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.
- (2006a). New perspectives on English sound patterns: ‘Natural’ and ‘unnatural’ in Evolutionary Phonology. *Journal of English Linguistics*, 34: 6–25.
- (2006b). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics*, 32: 117–66.
- (2008). Structure-preserving sound change: A look at unstressed vowel syncope in Austronesian. In *TBA*, A. Adelaar and A. Pawley (eds.). (to appear) Canberra: Pacific Linguistics.
- and Garrett, Andrew (1998). ‘The origins of consonant-vowel metathesis.’ *Language*, 74: 508–56.

- Bloomfield, Leonard (1895). On assimilation and adaptation in congeneric classes of words. *American Journal of Philology*, 16: 409–34.
- (1933). *Language*. Chicago: University of Chicago Press.
- Bochner, Harry (1993). *Simplicity in Generative Morphology*. Berlin: Mouton de Gruyter.
- Bod, Rens, Hay, Jennifer, and Jannedy, Stefanie (eds.) (2003). *Probabilistic Linguistics*. Cambridge: Cambridge University Press.
- Bohas, Georges, Guillaume, Jean-Patrick, and Kouloughli, Djamel Eddine (1990). *The Arabic Linguistic Tradition*. Arabic Thought and Culture. London: Routledge.
- Bonami, Olivier and Boyé, Gilles (2007). Remarques sur les bases de la conjugaison. In *Des sons et des sens*, E. Delais-Roussarie and L. Labruno (eds.). Paris: Hermès Sciences, 77–90.
- Braine, Martin D. S. (1966). Learning the positions of words relative to a marker element. *Journal of Experimental Psychology*, 72: 532–40.
- (1987). What is learned in acquiring word classes – a step toward an acquisition theory. In *Mechanisms of Language Acquisition*, B. MacWhinney (ed.). Hillsdale, NJ: Lawrence Erlbaum, 65–87.
- Brent, Michael R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34: 71–105.
- Broe, Michael (1993). *Specification Theory: The Treatment of Redundancy in Generative Phonology*. Ph.D. thesis, University of Edinburgh.
- Brooks, Patricia J., Braine, Martin D. S., Catalano, Lisa, Brody, Ruth E., and Sudhalter, Vicki (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, 32: 76–95.
- Buchholz, Eva (2004). *Grammatik der finnischen Sprache*. Bremen: Hempen Verlag.
- Burnard, Lou (1995). *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford: Oxford University Computing Service.
- Bybee, Joan L. (1985). *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.
- (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10: 425–55.
- (2001). *Phonology and Language Use*. Cambridge: Cambridge University Press.
- (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, 14: 261–90.
- (2006). From usage to grammar: The mind's response to repetition. *Language*, 82: 711–33.
- and Moder, Carol L. (1983). Morphological classes as natural categories. *Language*, 59: 251–70.
- and Pardo, Elly (1981). On lexical and morphological conditioning of alternations: A nonce-probe experiment with Spanish verbs. *Linguistics*, 19: 937–68.
- Cameron-Faulkner, Thea and Carstairs-McCarthy, Andrew (2000). Stem alternants as morphological signata: Evidence from blur avoidance in Polish nouns. *Natural Language and Linguistic Theory*, 18: 813–35.

- Campbell, Lyle (1998). *Historical Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- and Poser, William J. (2008). *Language Classification: History and Method*. Cambridge: Cambridge University Press.
- Caramazza, Alfonso (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14: 177–208.
- Carstairs, Andrew (1983). Paradigm economy. *Journal of Linguistics*, 19: 115–25.
- (1987). *Allomorphy in Inflection*. London: Croom Helm.
- Carstairs-McCarthy, Andrew (1991). ‘Inflection classes: Two questions with one answer’. In *Paradigms: The Economy of Inflection*, ed. F. Plank. Berlin: Mouton de Gruyter, 213–53.
- Casanto, Daniel and Boroditsky, Lera (2007). Time in the mind: Using space to think about time. *Cognition*, 102: 118–28.
- Chapman, Don and Royal Skousen (2005). Analogical modeling and morphological change: The case of the adjectival negative prefix in English. *English Language and Linguistics* 9(2): 1–25.
- Chater, Nick, Tenenbaum, Joshua B., and Yuille, Alan (eds.). (2006). Special issue: Probabilistic models of cognition. *Trends in Cognitive Sciences*, 10.7.
- Chitoran, Ioana and Hualde, José I. (2007). On the origin and evolution of the contrast between diphthongs and hiatus sequences in Romance. *Phonology*, 24: 37–75.
- Chomsky, Noam (1975). *The Logical Structure of Linguistic Theory*. Chicago: University of Chicago Press.
- (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.
- and Halle, Morris (1968). *The Sound Pattern of English*. New York: Harper and Row.
- and Lasnik, Howard (1977). Filters and control. *Linguistic Inquiry*, 8: 425–504.
- Clahsen, Harald, Aveledo, Fraibet, and Roca, Iggy (2002). The development of regular and irregular verb inflection in Spanish child language. *Journal of Child Language*, 29: 591–622.
- Sonnenstuhl, Ingrid, Hadler, Meike, and Eisenbeiss, Sonja (2001). Morphological paradigms in language processing and language disorders. *Transactions of the Philological Society*, 99: 247–77.
- Clark, Eve V. (1981). Lexical innovations: How children learn to create new words. In *The Child’s Construction of Language*, W. Deutsch (ed.). London: Academic Press, 299–328.
- (1983). Meaning and concepts. In *Handbook of Child Psychology: Vol. 3 Cognitive Development*, P. Mussen, L. Flavell, and E. Markman (eds.). New York: Wiley, 787–840.
- and Berman, Ruth A. (1987). Types of linguistic knowledge: Interpreting and producing compound nouns. *Journal of Child Language*, 14: 547–67.
- Gelman, Susan A., and Lane, Nancy M. (1985). Compound nouns and category structure in young children. *Child Development*, 56: 84–94.

- Costello, Fintan. J. and Keane, Mark T. (2001). Testing two theories of conceptual combination: Alignment versus diagnosticity in the comprehension and production of combined concepts. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27.1: 255–71.
- Cover, Thomas M. and Thomas, Joy A. (1991). *Elements of Information Theory*. New York: John Wiley & Sons.
- Crawley, Michael J. (2002). *Statistical Computing. An Introduction to Data Analysis using S-plus*. Chichester: Wiley.
- Croft, William (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- (2008). Evolutionary linguistics. *Annual Review of Anthropology*, 37: 219–34.
- Dabrowska, Ewa and Lieven, Elena (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, 16: 437–74.
- Daelemans, Walter and van den Bosch, Antal (2005). *Memory-based Language Processing*. Cambridge: Cambridge University Press.
- Zavrel, Jakob, van der Sloot, Ko, and van den Bosch, Antal (2000). TiMBL: Tilburg memory based learner reference guide 3.0. Technical report, Computational Linguistics Tilburg University.
- ——— ——— ——— (2005). TiMBL: Tilburg memory-based learner, version 5.1, reference guide. Technical report, ILK Technical Report Series 04–02.
- Daugherty, Kim and Seidenberg, Mark (1994). Beyond rules and exceptions: A connectionist approach to inflectional morphology. In *The Reality of Linguistic Rules*, S. D. Lima, R. L. Corrigan, and G. K. Iverson (eds.). Amsterdam: John Benjamins, 353–88.
- Deacon, Terrence W. (1997). *The Symbolic Species: The Co-Evolution of Language and the Human Brain*. London: Penguin Press.
- Deutscher, Guy (2001). On the mechanisms of morphological change. *Folia Linguistica Historica*, 22: 41–8.
- (2005). *The Unfolding of Language: The Evolution of Mankind's Greatest Invention*. London: Arrow.
- Di Sciullo, Anne-Marie and Williams, Edwin (1987). *On the Definition of Word*. Cambridge: MIT Press.
- Dinnsen, Daniel (1979). Maybe atomic phonology. In *Current Issues in Phonological Theory*, D. Dinnsen (ed.). Bloomington, IN: Indiana University Press, 31–49.
- Downing, Pamela A. (1977). On the creation and use of English compound nouns. *Language*, 53. 4: 810–42.
- Dressler, Wolfgang U., Libben, Gary, Stark, Jacqueline, Pons, Christiane, and Jarema, Gonia (2001). The processing of interfixed German compounds. In *Yearbook of Morphology 1999*, G. Booij and J. van Marle (eds.). Dordrecht: Kluwer, 185–220.
- Eddington, David (2002). A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling. In *Analogical modeling: An Exemplar-based Approach to Language*, R. Skousen, D. Lonsdale, and D. B. Parkinson (eds.). Amsterdam: John Benjamins, 141–55.

- Eisner, Frank and McQueen, James M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, 67: 224–38.
- Elzinga, Dirk (2006). English adjective comparison and analogy. *Lingua*, 116: 757–70.
- Erelt, Mati, Kasik, Reet, Metslang, Helle, Rajandi, Henno, Ross, Kristiina, Henn, Saari, Kaja, Tael, and Silvi, Vare (1995). *Eesti keele grammatika*. Volume I: Morfologia. Tallin: Eesti Teaduste Akadeemia Eesti Keele Instituut.
- Ernestus, Mirjam and Baayen, R. Harald (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79: 5–38.
- (2004). Analogical effects in regular past tense production in Dutch. *Linguistics*, 42: 873–903.
- Estes, Zachary (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language*, 48: 304–19.
- Fabb, Nigel (1998). Compounding. In *The Handbook of Morphology*, A. Spencer and A. M. Zwicky (eds.). Oxford: Blackwell Publishers, 66–83.
- Filipović Đurđević, Dušica (2007). *The Polysemy Effect in the Processing of Serbian Nouns*. Ph.D. thesis, University of Belgrade, Serbia.
- Finkel, Rafael and Stump, Gregory (2007). Principal parts and linguistic typology. *Morphology*, 17: 39–75.
- Firth, J. R. (1930). *Speech*. London: Ernest Benn.
- Friedman, Lynn and Wall, Melanie (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, 59: 127–36.
- Frigo, Lenore and McDonald, Janet L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, 39: 218–45.
- Frisch, Stefan A. (1996). *Similarity and Frequency in Phonology*. Ph.D. thesis, Northwestern University.
- Pierrehumbert, Janet B., and Broe, Michael B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, 22: 179–228.
- Gagné, Christina L. (2001). Relation and lexical priming during the interpretation of noun-noun combinations. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27: 236–54.
- and Shoben, Edward J. (1997). Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning Memory and Cognition*, 23: 71–87.
- (2002). Priming relations in ambiguous noun-noun combinations. *Memory and Cognition*, 30: 637–46.
- and Spalding, Thomas L. (2004). Effect of relation availability on the interpretation and access of familiar noun-noun compounds. *Brain and Language*, 90: 478–86.
- (2006). Conceptual combinations: Implications for the mental lexicon. In *The Representation and Processing of Compound Words*, G. Libben and G. Jarema (eds.). Oxford: Oxford University Press, 145–68.
- Gahl, Susanne and Yu, Alan C. L. (eds.). (2006). Special issue on exemplar-based models in linguistics. *The Linguistic Review*, 23.

- Garrett, Andrew (2008). Paradigmatic uniformity and markedness. See Good (2008), 125–43.
- Gentner, Dedre (1983). Structure mapping: a theoretical framework for analogy. *Cognitive Science*, 7: 155–70.
- Bowdle, Brian F., Wolff, Phillip, and Boronat, Consuelo (2001a). Metaphor is like analogy. See Gentner, Holyoak, and Kokinov (2001b), 199–253.
- Gentner, Dedre, Holyoak, Keith J., and Kokinov, Boicho N. (eds) (2001b). *The Analogical Mind: Perspectives from Cognitive Science*. Cambridge, MA: MIT Press.
- and Kurtz, Kenneth J. (2005). Relational categories. In *Categorization Inside and Outside the Lab*, W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, and P. W. Wolff (eds.). Washington, DC: APA, 151–75.
- and Markman, Arthur B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52: 45–56.
- Gerken, LouAnn, Wilson, Rachel, and Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language*, 32: 249–68.
- Gerken, LouAnn., Gómez, Rebecca L., and Nurmsoo, Erika (1999). The role of meaning and form in the formation of syntactic categories. Paper presented at the *Society for Research in Child Development*. Albuquerque, NM.
- Giegerich, Heinz J. (2004). Compound or phrase? English noun-plus-noun constructions and the stress criterion. *English Language and Linguistics*, 8: 1–24.
- Gildea, Daniel and Jurafsky, Daniel (1996). Learning bias and phonological-rule induction. *Computational Linguistics*, 22: 497–530.
- Gleitman, Lila R. and Gleitman, Henry (1970). *Phrase and Paraphrase: Some Innovative Uses of Language*. New York: W. W. Norton.
- Goldberg, Adele E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- and Jackendoff, Ray (2005). The end result(ative). *Language*, 81: 474–7.
- Goldinger, Stephen D. (2000). The role of perceptual episodes in lexical processing. In *Proceedings of SWAP, Workshop on Spoken Word Access Processes*, A. Cutler, J. M. McQueen, and R. Zondervan (eds.). Nijmegen: Max Planck Insititute for Psycholinguistics, 155–8.
- Goldsmith, John A. (2000). Linguistica: An automatic morphological analyzer. In *Papers from the 36th Annual Meeting of the Chicago Linguistic Society, Main Session*, A. Okrent and J. Boyle (eds.). Chicago: Chicago Linguistics Society, 125–39.
- (2001). The unsupervised learning of natural language morphology. *Computational Linguistics*, 27: 153–98.
- (2005). Review of Nevins. *Language*, 81: 719–36.
- (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12: 353–71.
- (2007). Towards a new empiricism. In *Recherches Linguistiques à Vincennes 36*, J. Brandao de Carvalho (ed.), 9–36.
- (2009). Segmentation and morphology. In *The Handbook of Computational Linguistics*, A. Clark, C. Fox, and S. Lappin (eds.). Oxford: Blackwell.

- and Hu, Yu (2004). From signatures to finite state automata. Technical Report TR-2005-05, Department of Computer Science, University of Chicago.
- Gómez, Rebecca L. and LaKusta, Laura (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, 7: 567–80.
- Good, Jeff (ed.) (2008). *Linguistic Universals and Language Change*. Oxford: Oxford University Press.
- Goswami, Usha (2001). Analogical reasoning in children. See Gentner, Holyoak, and Kokinov (2001), 437–70.
- and Brown, Ann L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, 35: 69–95.
- — (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition*, 36: 207–26.
- Grimshaw, Jane (1981). Form, function, and the language acquisition device. In *The Logical Problem of Language Acquisition*, C. L. Baker and J. J. McCarthy (eds.). Cambridge, MA: MIT Press, 165–82.
- Guenther, Frank H., Nieto-Castanon, Alfonso, Ghosh Satrajit, S., and Tourville, Jason A. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, and Hearing Research*, 47: 46–57.
- Gurevich, Olga (2006). *Construction Morphology: the Georgian Case*. Ph.D. thesis, University of California, Berkeley.
- Hafer, Margaret A. and Weiss, Stephen F. (1974). Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10: 371–85.
- Hahn, Ulricke and Chater, Nick (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? *Cognition*, 65: 197–230.
- — and Richardson, Lucy B. (2003). Similarity as transformation. *Cognition*, 87: 1–32.
- Halford, Graeme S. and Andrews, Glenda (2007). Domain general processes in higher cognition: Analogical reasoning, schema induction and capacity limitations. In *Integrating the Mind: Domain General versus Domain Specific Processes in Higher Cognition*, M. J. Roberts (ed.). New York: Psychology Press, 213–32.
- Halle, Morris (1962). Phonology in generative grammar. *Word*, 18: 54–72.
- and Marantz, Alec (1993). Distributed Morphology and the pieces of inflection. In *The View from Building 20*, K. Hale and S. J. Keyser (eds.). Cambridge, MA: MIT Press, 111–76.
- Hamp, Eric, Householder, Fred W., and Austerlitz, Robert (eds.) (1966). *Readings in Linguistics II*. Chicago: University of Chicago Press.
- Hare, Mary and Elman, Jeffrey L. (1995). Learning and morphological change. *Cognition*, 56: 61–98.
- Harrington, Jonathan, Palethorpe, Sallyanne, and Watson, Catherine I. (2000). Does the Queen speak the Queen's English? *Nature*, 408: 927–8.
- Harris, Alice C. and Campbell, Lyle (1995). *Historical Syntax in Cross-Linguistic Perspective*. Cambridge Studies in Linguistics, vol. 74. Cambridge: Cambridge University Press.

- Harris, Zellig S. (1942). Morpheme alternants in linguistic analysis. *Language*, 18: 169–180. Reprinted in Joos (1957), 109–15.
- (1955). From phoneme to morpheme. *Language*, 31: 190–222.
- Harris, Zellig S. (1967). Morpheme boundaries within words: Report on a computer test. In *Papers in Structural and Transformational Linguistics (1970)*, 68–77. Dordrecht: D. Reidel.
- Hay, Jennifer (2001). Lexical frequency in morphology: Is everything relative? *Linguistics*, 39: 1041–70.
- Hay, Jennifer B. and Baayen, R. Harald (2002). Parsing and productivity. In *Yearbook of Morphology 2002*, Geert E. Booij and Jaapvan Marle (eds.). Dordrecht: Kluwer, 203–35.
- and — (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, 9: 342–8.
- Pierrehumbert, Janet B., and Beckman, Mary (2004). Speech perception, well-formedness and the statistics of the lexicon. In *Phonetic Interpretation: Papers in Laboratory Phonology VI*, J. Local, R. Ogden, and R. Temple (eds.). Cambridge: Cambridge University Press, 58–74.
- Heine, Bernd (1993). *Auxiliaries: Cognitive Forces and Grammaticalization*. Oxford: Oxford University Press.
- Hinton, Leanne, Nichols, Johanna, and Ohala, John J. (eds.). (1995). *Sound Symbolism*. Cambridge: Cambridge University Press.
- Hock, Hans Henrich (1991). *Principles of Historical Linguistics*. Berlin: Mouton de Gruyter.
- (2003). Analogical change. In *The Handbook of Historical Linguistics*, B. Josephs and R. Janda (eds.). Oxford: Oxford University Press, 441–80.
- Hockett, Charles F. (1947). Problems of morphemic analysis. *Language*, 23: 321–43. Reprinted in Joos (1957), 229–42.
- (1954). Two models of grammatical description. *Word*, 10: 210–31. Reprinted in Joos (1957), 386–99.
- (1960). The origin of speech. *Scientific American*, 203: 88–9.
- (1966). *Language, mathematics and linguistics*. In T. Seberk (ed.) *Current trends in linguistics, vol 3: Theoretical Foundations*. The Hague: Mouton. 155–304.
- (1968). *The State of the Art*. The Hague: Mouton.
- (1987). *Refurbishing our Foundations: Elementary Linguistics from an Advanced Point of View*. Current Issues in Linguistic Theory, vol. 56. Amsterdam: John Benjamins.
- Holyoak, Keith J. and Thagard, Paul (1977). The analogical mind. *American Psychologist*, 52: 35–44.
- and — (1989). Analogical mapping by constraint satisfaction. *Cognitive Science* 13: 295–355.
- and — (1997). The analogical mind. *American Psychologist*, 52(1): 35–44.
- Hopper, Paul and Traugott, Elizabeth C. (2003). *Grammaticalization*. Cambridge: Cambridge University Press.
- Hu, Yu, Matveeva, Irina, Goldsmith, John A., and Sprague, Colin (2005). Using morphology and syntax together in unsupervised learning. In *Proceedings of the*

- Workshop on Psychocomputational Models of Human Language Acquisition*, Ann Arbor, MI: Association for Computational Linguistics, 20–7.
- Huang, Hsu-Wen, Lee, Chia-Ying, Tsai, Jie-Li, Lee, Chia-Lin, Hung, Daisy L., and Tzeng, Ovid J.-L. (2006). Orthographic neighborhood effects in reading Chinese two-character words. *Neuroreport*, 17: 1061–5.
- Hughes, Michael and Ackerman, Farrell (2002). Words and paradigms: Estonian nominal declension. In *Papers from the 37th Annual Meeting of the Chicago Linguistics Society*, M. Andronis, C. Ball, H. Elston, and S. Neuvel (eds.).
- Hunt, Gavin R. and Gray, Russell D. (2004). The crafting of hook tools by wild New Caledonian crows. *Proceedings of the Royal Society of London, B. Biol. Sci.*, 271: S88–S90.
- Itkonen, Esa (2005). *Analogy as Structure and Process*. Amsterdam: John Benjamins.
- Jackendoff, Ray (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.
- Jakobi, Angelika (1990). *A Fur Grammar*. Hamburg: Helmut Buske Verlag.
- Johnson, C. Douglas (1972). *Formal Aspects of Phonological Description*. The Hague: Mouton.
- Johnson, Keith (1997). Speech perception without speaker normalization: An exemplar model. In *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix (eds.). San Diego: Academic Press, 145–65.
- de Jong, Nivja H. (2002). *Morphological Families in the Mental Lexicon*. Ph.D. thesis, University of Nijmegen.
- Feldman, Laurie B., Schreuder, Robert, Pastizzo, Matthew J., and Baayen, R. Harald (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, 81: 555–67.
- Schreuder, Robert, and Baayen, R. Harald (2000). The morphological family size effect and morphology. *Language and Cognitive Processes*, 15: 329–65.
- Joos, Martin (ed.) (1957). *Readings in Linguistics I*. Chicago: University of Chicago Press.
- Joseph, Brian and Janda, Richard (1988). The how and why of diachronic morphologization and demorphologization. In *Theoretical Morphology: Approaches in Modern Linguistics*, M. Hammond and M. Noonan (eds.). San Diego: Academic Press, 193–210.
- Kang, Yoonjung (2006). Neutralization and variations in Korean verbal paradigms. In *Harvard Studies in Korean Linguistics XI*. Hanshin Publishing Company, 183–96.
- Kaplan, Ronald M. and Kay, Martin (1994). Regular models of phonological rule systems. *Computational Linguistics*, 20: 331–78.
- Karlsson, Fred (1999). *Finnish: An Essential Grammar*. London: Routledge.
- Katz, Jeffrey S. and Wright, Anthony A. (2006). Same/different abstract-concept learning by pigeons. *Journal of Experimental Psychology: Animal Behavior Proceedings*, 32: 80–6.
- Kay, Paul and Fillmore, Charles J. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language*, 75: 1–33.

- and Zimmer, Karl (1976). On the semantics of compounds and genitives in English. In *Sixth California Linguistics Association Proceedings*, San Diego, CA: Campile Press, 29–35.
- Kazazis, Kostas (1969). Possible evidence for (near-)underlying forms in the speech of a child. *Chicago Linguistics Society*, 5: 382–6.
- Kemler Nelson, Deborah, Jusczyk, Peter W., Mandel, Denise R., Myers, James, Turk, Alice E., and Gerken, LouAnn (1995). The headturn preference procedure for testing auditory perception. *Infant Behavior and Development*, 18: 111–16.
- Kempe, Vera and Brooks, Patricia J. (2001). The role of diminutives in the acquisition of Russian gender: Can elements of child-directed speech aid in learning morphology? *Language Learning*, 51: 221–56.
- Keuleers, Emmanuel (2008). *Memory-based Learning of Inflectional Morphology*. Antwerp: University of Antwerp.
- Sandra, Dorniniek, Daelemans, Walter, Gillis, Steven, Durieux, Gert, and Martens, Evelyn (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology*, 54: 283–318.
- Kibrik, Aleksandr E. (1998). Archi. In *Handbook of Morphology*, A. Spencer and A. M. Zwicky (eds.). Oxford: Blackwell, 455–76.
- Kirby, Simon (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5: 102–10.
- (2007). The evolution of language. In *Oxford Handbook of Evolutionary Psychology*, R. Dunbar and L. Barrett (eds.). Oxford: Oxford University Press, 669–81.
- Kostić, Aleksandar (1991). Informational approach to processing inflected morphology: Standard data reconsidered. *Psychological Research*, 53: 62–70.
- (1995). Informational load constraints on processing inflected morphology. In *Morphological Aspects of Language Processing*, L. B. Feldman (eds.). Hillsdale, NJ: Lawrence Erlbaum, 317–44.
- (2008). The effect of the amount of information on language processing. In submission.
- Marković, Tanja, and Baucal, Aleksandar (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative domains? In *Morphological Structure in Language Processing*, R. H. Baayen and R. Schreuder (eds.). Berlin: Mouton de Gruyter, 1–44.
- Kraska-Szlenk, Iwona (2007). *Analogy: The Relation between Lexicon and Grammar*. Munich: Lincom Europa.
- Krott, Andrea, Gagné, Christina L., and Nicoladis, Elena (2009). How the parts relate to the whole: Frequency effects on childrens’s interpretation of novel compounds. *Journal of Child Language*, 36: 85–112.
- Hagoort, Peter, and Baayen, R. Harald (2004). Sublexical units and supralexicall combinatorics in the processing of interfixed Dutch compounds. *Language and Cognitive Processes*, 19: 453–71.
- Krebbers, Loes, Schreuder, Robert, and Baayen, R. Harald (2002a). Semantic influence on linkers in Dutch noun-noun compounds. *Folia Linguistica*, 36: 7–22.

- and Nicoladis, Elena (2005). Large constituent families help children parse compounds. *Journal of Child Language*, 32: 139–58.
- Schreuder, Robert, and Baayen, R. Harald (2001). Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics*, 1: 51–93.
- — — (2002*b*). Analogical hierarchy: Exemplar-based modeling of linkers in Dutch noun-noun compounds. In *Analogical Modeling: An Exemplar-Based Approach to Language*, R. Skousen, D. Londsedale, and D. B. Parkinson (eds.). Amsterdam: John Benjamins, 181–206.
- — — (2002*c*). Linking elements in Dutch noun-noun compounds: Constituent families as analogical predictors for response latencies. *Brain and Language*, 81: 708–22.
- — — and Dressler, Wolfgang U. (2007). Analogical effects on linking elements in German compounds. *Language and Cognitive Processes*, 22: 25–57.
- Kruskal, Joseph B. (1983). An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and J. B. Kruskal (eds.). Reading, MA: Addison-Wesley, 1–44.
- Kuehne, Sven E., Forbus, Kenneth D., Gentner, Dedre, and Quinn, Bryan (2000). SEQL: Category learning as progressive abstraction using structure mapping. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, 770–5.
- Kuhl, Patricia K. (1991). Human adults and human infants show a perceptual magnet effect for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50: 93–107.
- (1995). Mechanisms of developmental change in speech and language. In *Proceedings of the XIIIth International Congress of Phonetic Sciences, Vol. 2*, K. Elenius and P. Branderud (eds.). Stockholm: Stockholm University, 132–9.
- Kuperman, Victor, Bertram, Raymond, and Baayen, R. Harald (2008). Morphological dynamics in compound processing. In submission.
- Kupryanova, Z. N. (1985). *Nenetskij jazyk [Nenets Language]*. Moscow: Nauk.
- Kuryłowicz, Jerzy (1947). La nature des procès dits “analogiques”. *Acta Linguistica*, 5: 121–38. Reprinted in Hamp, Householder, and Austerlitz (1966), 158–74. English translation with introduction by Margaret Winters (1995), The nature of the so-called analogical processes. *Diachronica* 12: 113–45.
- Lahiri, Aditi (ed.) (2000). *Analogy, Levelling, Markedness: Principles of Change in Phonology and Morphology*. Berlin: Mouton de Gruyter.
- Lakoff, George and Johnson, Mark (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
- Landauer, Thomas K. and Dumais, Susan T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104: 211–40.
- Law, Vivien (2003). *The History of Linguistics in Europe: From Plato to 1600*. Cambridge: Cambridge University Press.
- Lees, Robert B. (1960). *The Grammar of English Nominalizations*. The Hague: Mouton de Gruyter.

- Levelt, Willem J. M., Roelofs, Ardi, and Meyer, Antje S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22: 1–37.
- Levi, Judith N. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.
- Levy, Roger (2008). Expectation-based syntactic comprehension. *Cognition*, 106: 1126–77.
- Libben, Gary, Jarema, Gonia, Dressler, Wolfgang, Stark, Jacqueline, and Pons, Christiane (2002). Triangulating the effects of interfixation in the processing of German compounds. *Folia Linguistica*, 36: 23–43.
- Lieber, Rochelle (1992). *Deconstructing Morphology*. Chicago: University of Chicago Press.
- Lieven, Elena, Behrens, Heike, Speares, Jennifer, and Tomasello, Michael (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30: 333–70.
- Locke, John (1690/1975). *An Essay Concerning Human Understanding*. History of Economic Thought Books. McMaster University Archive for the History of Economic Thought, Canada: Hamilton.
- Long, Christopher J. and Almor, Amit (2000). Irregularization: The interaction of item frequency and phonological interference in regular past tense production. In *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, 310–15.
- Lonsdale, Deryle (2002). Data files for analogical modeling. In *Analogical Modeling: An Exemplar-Based Approach to Language*, Royal Skousen, Deryle Lonsdale, and Dilworth B. Parkinson (eds.). 349–63. Amsterdam: John Benjamins.
- Luce, R. Duncan (1959). *Individual Choice Behavior*. New York: Wiley.
- MacWhinney, Brian. (2000). *The Childes Project: Tool for Analyzing Talk, Vol. 1: Transcription Format and Programs. Vol. 2: The Database*. Mahwah, NJ: Lawrence Erlbaum.
- MacWhinney, Brian and Leinbach, Jared (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, 40: 121–57.
- Magnus, Margaret (2000). *What's in a Word? Evidence for Phonosemantics*. Ph.D. thesis, University of Trondheim.
- Mańczak, Witold (1958). Tendances générales des changements analogiques. *Lingua*, 7: 298–325, 387–420.
- (1980). Laws of analogy. In *Historical Morphology* J. Fisiak (ed.). The Hague: Mouton, 283–8.
- Manning, Christopher D. and Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.
- Maratsos, Michael (1982). The child's construction of grammatical categories. In *Language Acquisition: The State of the Art*, E. Wanner and L. Gleitman (eds.). Cambridge: Cambridge University Press.
- de Marcken, Carl (1996). *Unsupervised Language Acquisition*. Ph.D. thesis, MIT, Cambridge MA.
- Marcus, Gary F., Brinkmann, Ursula, Clahsen, Harald, Wiese, Richard, and Pinker, Steven (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, 29: 189–256.

- Mattens, W. H. M. (1984). De voorspelbaarheid van tussenklanken in nominale samenstellingen. *De Nieuwe Taalgids*, 7: 333–43.
- Matthews, Peter H. (1972). *Inflectional Morphology: A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge: Cambridge University Press.
- (1974). *Morphology. An Introduction to the Theory of Word Structure*. Cambridge: Cambridge University Press.
- (1991). *Morphology*. Cambridge: Cambridge University Press.
- (2007). *Syntactic Relations: A Critical Survey*. Vol. 114, Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.
- Mayr, Ernst (1997). The objects of selection. *Proceedings of the National Academy of Sciences*, 94: 2091–4.
- McClelland, James L. and Patterson, Karalyn (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, 6: 465–72.
- Meillet, Antoine (1915). *Étude comparative des langues indo-européennes*. Paris: Hachette et Cie.
- Mellenius, Ingmarie (1997). *The Acquisition of Nominal Compounding in Swedish*. Lund: Lund University Press.
- Merlo, Lauren M. F., Pepper, John W., Reid, Brian J., and Maley, Carlo C. (2006). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer*, 6: 924–35.
- Merrifield, William R. and Rensch, Calvin R. (eds.) (1990). *Syllables, Tone, and Verb Paradigms*. Studies in Chinantec Languages, Vol. 4. Summer Institute of Linguistics and The University of Texas at Arlington, Dallas.
- Mielke, Jeff (2004). *The Emergence of Distinctive Features*. Ph.D. thesis, Ohio State University.
- (2008). *The Emergence of Distinctive Features*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.
- Milin, Petar, Filipović Đurđević, Dušica, and Moscogo del Prado Martín, Fermín (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, 60: 50–64.
- Miller, George A. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3: 235–312.
- Mintz, Tobin (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition*, 30: 678–86.
- (2006). Frequent frames: Simple co-occurrence constructions and their links to linguistic structure. In *Constructions in Acquisition*, B. Kelly and E. V. Clark (eds.). Stanford: CSLI, 59–82.
- Miozzo, Michele and Caramazza, Alfonso (2005). The representation of homophones: Evidence from the distractor-frequency effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31: 1360–71.
- Morpurgo Davies, Anna (1978). Analogy, segmentation and the early Neogrammarians. *Transactions of the Philological Society*, 36–60.
- Morris, Richard E (2005). Attraction to the unmarked in Old Spanish leveling. In *Selected Proceedings of the 7th Hispanic Linguistics Symposium*, D. Eddington (ed.). Somerville, MA: Cascadilla Proceedings Project, 180–91.

- Moscoso del Prado Martín, Fermín (2003). *Paradigmatic Structures in Morphological Processing: Computational and Cross-Linguistic Studies*. Ph.D. thesis, University of Nijmegen.
- Bertram, Raymond, Haikio, Tuomo, Schreuder, Robert, and Baayen, R. Harald (2004a). Morphological family size in a morphologically rich language: The case of Finnish compared to Dutch and Hebrew. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 30: 1271–8.
- Kostić, Aleksandar, and Filipović Đurđević, Dušica (2008). The missing link between morphemic assemblies and behavioral responses: A Bayesian information-theoretical model of lexical processing. Manuscript submitted for publication.
- ——— and Baayen, R. Harald (2004b). Putting the bits together: An information-theoretical perspective on morphological processing. *Cognition*, 94: 1–18.
- Murphy, Gregory L. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language*, 29: 259–88.
- Nakisa, Ramin Charles, Plunkett, Kim, and Hahn, Ulrike (2000). A cross-linguistic comparison of single and dual-route models of inflectional morphology. In *Cognitive Models of Language Acquisition*, P. Broeder and J. Murre (eds.). Cambridge, MA: MIT Press, 201–22.
- Neuvel, Sylvain and Fulop, Sean A. (2002). Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, Morristown, NJ: Association for Computational Linguistics, 31–40.
- and Singh, Rajendra (2001). Vive la différence! What morphology is about. *Folia Linguistica*, 35: 313–20.
- New, Boris, Brysbaert, Marc, Segui, Juan, Ferrand, Ludovic, and Rastle, Kathleen (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language*, 51: 568–85.
- Nickel, Klaus Peter (1990). *Samisk grammatikk*. Oslo: Universitetsforlaget.
- Nicoladis, Elena (2003). What compound nouns mean to preschool children. *Brain and Language*, 84: 38–49.
- and Krott, Andrea (2007). Family size and French-speaking children's segmentation of existing compounds. *Language Learning*, 57: 201–228.
- Nosofsky, Robert M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115: 39–57.
- (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology*, 34: 393–418.
- Oinas, Zsuzsanna (2008). *Guide to Finnish Declension*. Espoo: Finnlibri.
- Pace, Wanda J. (1990). Comaltepec Chinantec verb inflection. See Merrifield and Rensch (1990), 21–62.
- Parault, Susan J., Schwanenflugel, Paula J., and Haverback, Heather R. (2005). The development of interpretations for novel noun-noun conceptual combinations during the early school years. *Journal of Experimental Child Psychology*, 91: 67–87.
- Paul, Hermann (1920). *Prinzipien der Sprachgeschichte*. Tübingen: Max Niemayer Verlag.

- Penn, Derek C., Holyoak, Keith J., and Povinelli, Daniel J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, 31: 109–78.
- Penny, Ralph (2002). *A History of the Spanish Language*. Cambridge: Cambridge University Press.
- Pepperberg, Irene M. (1987). Acquisition of the same/different concept by an African Grey parrot (*Psittacus erithacus*): Learning with respect to categories of color, shape and material. *Animal Learning and Behavior*, 15: 423–32.
- Pierrehumbert, Janet B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In *Frequency and the Emergence of Linguistic Structure*, J. Bybee and P. Hopper (eds.). Amsterdam: John Benjamins, 137–58.
- (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, 46: 115–54.
- (2006). The new toolkit. *Journal of Phonetics*, 34: 516–30.
- Pihel, Kalju and Pikamäe, Arno (eds.) (1999). *Soome-eeesti sõnaraamat*. Tallinn: Valgus.
- Pinker, Steven (1982). A theory of the acquisition of lexical interpretive grammars. In *The Mental Representation of Grammatical Relations*, J. Bresnan (ed.). Cambridge, MA: MIT Press, 655–726.
- (1991). Rules of language. *Science*, 153: 530–35.
- (1999). *Words and Rules: The Ingredients of Language*. London: Weidenfeld and Nicolson.
- and Ullman, Michael T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, 6: 456–63.
- Plag, Ingo (2006). The variability of compound stress in English: Structural, semantic, and analogical factors. Part 1. *English Language & Linguistics*, 10: 143–72.
- Premack, David (1983). Animal cognition. *Annual Review of Psychology*, 34: 351–62.
- Rissanen, Jorma (1989). *Stochastic Complexity in Statistical Inquiry*. Riner Edge, NJ: World Scientific Publishing Company.
- Roark, Brian and Sproat, Richard (2007). *Computational Approaches to Morphology and Syntax*. Oxford: Oxford University Press.
- Robins, Robert H. (1959). In defense of WP. *Transactions of the Philological Society*, 116–44. Reprinted in *Transactions of the Philological Society* 99: 1–36.
- Robinson, A. H. (1979). Observations on some deficiencies in the transformational model as applied to particular compound types in French. *Cahiers de Lexicologie*, 35: 107–15.
- Ross, Sheldon M. (1988). *A First Course in Probability*. New York: Macmillan Publishing Company.
- Rubach, Jerzy and Booij, Geert (1985). A grid theory of Polish stress. *Lingua*, 66: 281–319.
- Rumelhart, David E. and McClelland, James L. (1987). Learning the past tenses of English verbs. In *Mechanisms of Language Acquisition*, B. MacWhinney (ed.). Hillsdale, NJ: Lawrence Erlbaum, 194–248.
- Ryder, Mary Ellen (1994). *Ordered Chaos: An Investigation of the Interpretation of English Noun-Noun Compounds*. Berkeley and Los Angeles: University of California Press.

- Salminen, Tapani (1993). On identifying basic vowel distinctions in Tundra Nenets. *Finno-Ugrische Forschungen*, 51: 177–87.
- (1997). *Tundra Nenets Inflection*. Mémoires de la Société Finno-Ougrienne 227, Helsinki.
- (1998). *A Morphological Dictionary of Tundra Nenets*. Lexica Societatis Fenno-Ugricae 26, Helsinki.
- Sankoff, Gillian and Blondeau, Hélène (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 83: 560–88.
- Sapir, Edward (1921). *Language*. New York: Harcourt Brace.
- Saussure, Ferdinand de (1916). *Cours de linguistique générale*. Paris: Payot.
- Schreuder, Robert and Baayen, R. Harald (1995). Modeling morphological processing. In *Morphological Aspects of Language Processing*, L. B. Feldman (ed.). Hillsdale, NJ: Lawrence Erlbaum, 131–54.
- (1997). How complex simplex words can be. *Journal of Memory and Language*, 37: 118–39.
- Seiler, Hansjakob (1965). On paradigmatic and syntagmatic similarity. *Lingua*, 18: 35–97.
- Shannon, Claude (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379–423, 623–56.
- Sharkey, William S. (1982). *The Theory of Natural Monopoly*. Cambridge: Cambridge University Press.
- Skousen, Royal (1989). *Analogical Modeling of Language*. Dordrecht: Kluwer.
- (1992). *Analogy and Structure*. Dordrecht: Kluwer.
- (2002). Analogical modeling and quantum computing. In *Analogical Modeling*, Skousen, Royal, Deryle Lonsdale, and Dilworth B. Parkinson (eds.), 319–46.
- (2003) Analogical Modeling: Exemplars, Rules, and Quantum Computing. *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society*, Pawel Nowak, Corey Yoquelet, and David Mortensen (eds.), 425–39 [also available at <<http://humanities.byu.edu/am/>>].
- (2005). Quantum Analogical Modeling: A General Quantum Computing Algorithm for Predicting Language Behavior. Preprint, posted under Quantum Physics on <arXiv.org>, quant-ph/0510146, October 18, 2005.
- Skousen, Royal, Lonsdale, Deryle, and Parkinson, Dilworth B. (eds.) (2002). *Analogical Modeling: An Exemplar-Based Approach to Language*. Amsterdam: John Benjamins.
- Smith, Kirk H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, 72: 580–8.
- Sproat, Richard (2008). Experiments in morphological evolution. Keynote address. 3rd Workshop on Quantitative Investigations in Theoretical Linguistics, Helsinki.
- Stampe, David (1980). *How I Spent My Summer Vacation*. New York: Garland Press.
- Stemberger, Joseph P. and MacWhinney, Brian (1986). Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition*, 14: 17–26.
- Steriade, Donca (2000). Paradigm uniformity and the phonetics-phonology boundary. In *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, M. B. Broe and J. B. Pierrehumbert (eds.). Cambridge: Cambridge University Press, 313–34.

- Storms, Gert and Wisniewski, Edward J. (2005). Does the order of head noun and modifier explain response times in conceptual combination? *Memory and Cognition*, 33: 852–61.
- Stump, Gregory T. (2001). *Inflectional Morphology: A Theory of Paradigm Structure*. Cambridge: Cambridge University Press.
- Sturtevant, Edgar H. (1947). *An Introduction to Linguistic Science*. New Haven: Yale University Press.
- Taylor, Alex H., Hunt, Gavin R., and Holzhaider, Jennifer C. Gray, Russell D. (2007). Spontaneous metatool use by New Caledonian crows. *Current Biology*, 17: 1504–7.
- Taylor, Douglas R, Zeyl, Clifford, and Cooke, Erin (2002). Conflicting levels of selection in the accumulation of mitochondrial defects in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 99: 3690–4.
- Tenenbaum, Joshua B. and Griffiths, Thomas L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24: 629–40.
- Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin and Review*, 2: 339–63.
- Tereshchenko, Natal'ya Mitrofanovna (1965). *Nenetsko-russkij slovar' [Nenets-Russian dictionary]*. St. Petersburg: Sovetskaja Entsiklopedija.
- Thymé, Anne E. (1993). *A Connectionist Approach to Nominal Inflection: Paradigm Patterning and Analogy in Finnish*. Ph.D. thesis, University of California, San Diego.
- Ackerman, Farrell, and Elman, Jeffrey L. (1994). Finnish nominal inflections: Paradigmatic patterns and token analogy. In *The Reality of Linguistic Rules*, S. D. Lima, R. L. Corrigan, and G. K. Iverson (eds.). Amsterdam: John Benjamins, 445–66.
- Tomasello, Michael (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11: 61–82.
- (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.
- Traficante, Daniela and Burani, Cristina (2003). Visual processing of Italian verbs and adjectives: The role of inflectional family size. In *Morphological Structure in Language Processing*, R. H. Baayen and R. Schreuder (eds.). Berlin: Mouton de Gruyter, 45–64.
- Traugott, Elizabeth C. and Heine, Bernd (1991). *Approaches to Grammaticalization*. Vol. 19, Typological Studies in Language. Amsterdam: John Benjamins.
- Tsai, Jie-Li, Lee, Chia-Ying, Lin, Ying-Chun, Tzeng, Ovid J.-L., and Hung, Daisy L. (2006). Neighborhood size effects of Chinese words in lexical decision and reading. *Language and Linguistics*, 7. 3: 659–75.
- Tschenkéli, Kita (1958). *Einführung in die georgische Sprache*. Zurich: Amirani Verlag.
- Tversky, Amos (1977). Features of similarity. *Psychological Review*, 84: 327–52.
- Vance, Timothy J. (1980). The psychological status of a constraint on Japanese consonant alternation. *Linguistics*, 18: 145–67.
- van den Toorn, M. C. (1982a). Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen I. *De Nieuwe Taalgids*, 75: 24–33.
- (1982b). Tendenzen bij de beregeling van de verbindingsklank in nominale samenstellingen II. *De Nieuwe Taalgids*, 75: 153–60.

- van Jaarsveld, Henk J., Coolen, Riet, and Schreuder, Robert (1994). The role of analogy in the interpretation of novel compounds. *Journal of Psycholinguistic Research*, 23: 111–37.
- Vennemann, Theo H. (1972). Phonetic analogy and conceptual analogy. In *Schuchardt, the Neogrammarians, and the Transformational Theory of Phonological Change: Four Essays*, T. Vennemann and T. H. Wilbur (eds.). Frankfurt: Athenaeum, 181–204.
- (1993). Language change as language improvement. In *Historical Linguistics: Problems and Perspectives*, C. Jones (ed.). London: Longman, 319–44.
- Viks, Ülle (1992). *Väike vormi-sõnastik: Sissejuhatus & grammatika*. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.
- Wedel, Andrew (2004). *Self-Organization and Categorical Behavior in Phonology*. Ph.D. thesis, University of California, Santa Cruz.
- (2006). Exemplar models, evolution and language change. *The Linguistic Review*, 23: 247–74.
- (2007). Feedback and regularity in the lexicon. *Phonology*, 24: 147–85.
- Wheeler, Max W. (2005). *The Phonology of Catalán*. Oxford: Oxford University Press.
- Whitney, William D. (1875). *The Life and Growth of Language*. London: H. S. King.
- Wilson, Rachel (2002). *Syntactic Category Learning in a Second Language*. Ph.D. thesis, University of Arizona, Tucson.
- Wisniewski, Edward J. (1996). Construal and similarity in conceptual combination. *Journal of Memory and Language*, 35: 434–53.
- Wurzel, Wolfgang U. (1970). *Studien zur deutschen Lautstruktur*. Berlin: Akademie-Verlag.
- (1989). *Natural Morphology and Naturalness*. Dordrecht: Kluwer.
- Yu, Alan C. L. (2007). Tonal phonetic analogy. Paper presented at ICPHS, Saarbrücken.
- Zanone, P. G. and Kelso, J. A. S. (1997). The coordination dynamics of learning and transfer: Collective and component levels. *Journal of Experimental Psychology, Human Perception and Performance*, 23: 1454–80.
- Zwicky, Arnold M. (1985). How to describe inflection. In *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, M. Niepokuj, M. Van Clay, V. Nikiporidou, and D. Feder (eds.). Berkeley: Berkeley Linguistics Society, 372–86.

Index

NAMES

Albright, Adam 2, 6, 12, 50, 56f., 61, 74, 8, 89–91, 95, 190–193, 195, 198, 204–206, 209f., 251f.
Anttila, Raimo 2, 4
Baayen, Harald 3, 6, 10, 61, 81, 90f., 119–125, 127, 135, 153, 184, 205, 214f., 221, 227–229, 232f., 235–238, 247, 250f.
Bennett, Charles 167
Bloomfield, Leonard 2, 6f., 58, 119
Bourgerie, Dana 171–173
Bybee, Joan 10, 87, 90f., 118, 184, 205–209
Chomsky, Noam 9, 83, 138, 142, 146
Eddington, David 101, 114, 181
Elzinga, Dirk 114, 182
Garrett, Andrew 2, 6
Gentner, Dedre 1, 5, 7, 90, 101, 114, 185
Harris, Zellig 55, 148, 155
Hockett, Charles F. 2, 7, 55, 61, 215
Holyoak, Keith J. 1, 3, 5, 114, 185
Itkonen, Esa 2, 9, 84, 99
Kostić, Aleksandar 224–226, 228, 232, 249
Kuryłowicz, Jerzy 6, 209
Locke, John 137, 162
Mańczak, Witold 6, 209
Matthews, Peter H. 10, 14, 57, 60f., 215, 235

Meillet, Antoine 151
Paul, Hermann 2, 5
Pinker, Steven 10, 90, 102, 118, 214
Sapir, Edward 54
Saussure, Ferdinand de 54, 59
Skousen, Royal 2, 6, 11, 95, 114–117, 122, 165, 167–171, 174–176
Tomasello, Michael 8, 132, 134
SUBJECTS
acquisition 8, 11, 18, 88, 102, 118, 120, 130–134, 162
algorithm 11, 18, 95, 122, 137, 139, 146–148, 153, 155, 157f., 161, 210
amplitude 166
Analogical Modeling 6, 11, 95, 115, 122, 164, 166, 168, 174–175, 185
analogical set 174–176, 187–190, 200–202
analogy
 brief history of 2, 13
 directionality of 6, 8, 186, 208–212
 analogy, four-part (see also analogy, proportional) 2, 3, 5, 187–189, 197, 251f.
 in child language 4, 8, 136
 in compound words 4f., 11, 118–121, 123–124, 126, 129, 133, 136
 in language change 2, 4, 5, 6, 8, 84, 86, 100, 185f., 188, 190, 208, 212
 in lexical category learning 101,
 morphological 5, chapter 7
 phonetic 6

- analogy (*Cont.*)
 phonological 7
 proportional (see also analogy, four-part) 2, 4, 11, 57, 62, 197, 245f.
 semantic 6–8, 129, 242
 syntactic 8f.,
 traditional 9f., 62, 137, 139, 156f., 165f., 187, 251
 type-frequency driven 205–207, 212, 251f.
 word-based 3–6, 9, 189–192, 197, 204
- associative relations 59
- base words 217, 229, 237–238, 242f., 245, 247
- Bayes' rule 114, 142f., 220
- boundaries 134, 139–141, 143, 146, 155, 158
- Brown Corpus 146f., 152
- category
 formation 102, 104, 107, 114, 116f.
 learning 101f., 105, 110, 112f., 115–117
 relational 90
- categorization 2, 85, 104, 106, 131, 164
 errors in 85
 similarity bias in 85
- CELEX database 120, 123, 232, 236f.
- cell predictability 11, 14, 36, 39, 41, 44, 50–52, 57, 63, 67–69, 71, 80
- Chinese classifiers 184
- cognition 1, 11
- competition 91f., 95, 97, 100, 122, 128f.
- compounds 4f., 118–121, 123–136, 235f.
 acquisition of 120, 130–134
 interpretation of 120, 128–135
 production of 120, 125–127, 130, 134f.
 visual processing of 118, 126f., 129f., 135f.
- computation 149, 155, 210
- computational models 95, 117, 122, 124, 138, 166, 178, 185f., 190, 193, 212f., 251
- computational problems 117, 138–140, 166, 178, 186
- connectionist network 89, 118, 122
- connectivity 124f., 216, 232, 235f., 248, 151
- constituent families 119, 121–128, 130, 132f., 135, 232, 236
- construction grammar 8f., 58f.
- context-sensitive rules 12, 186, 192f., 195, 205, 211f.
- core theory 164, 167
- cues 11, 79, 81, 102, 104–117, 140
 distributional 106, 117
 to lexical category formation 102, 107, 116f.
- dataset 164, 170, 172f., 179, 184, 195, 198, 243, 251
- derivation 3, 5, 73–75, 108, 205, 214f., 224, 229, 233, 235–238, 248, 252
- derived words 4, 73, 75, 80, 211, 215f., 236–247, 251
- directionality 93, 208, 212
- distinctive features 174, 181
- English adjective comparison 153, 179, 182
- English indefinite article 177, 179
- English negative prefix 179, 182
- English possessives 175
- entropy 52, 62–65, 67f., 70f., 75–78, 221, 223–235, 237–248, 250–252
 conditional 52, 6267f., 70f., 75, 77f
 cross 239–248, 251f.
 expected conditional 62
 relative 223f., 234f., 237–241, 243, 248
- errors 86, 89, 91, 146, 177, 182, 185, 203, 206, 208, 211f., 215, 234
- evolution 5, 10, 88, 91–94, 99, 152
 biological 91f., 99
 linguistic 5, 10
- evolutionary framework 84f., 88, 93
- Evolutionary Phonology 87, 91
- exemplar 9, 11, 57, 89, 91, 118, 122–124, 126, 132, 164–166, 175f., 179, 184, 191, 205, 207, 235, 251f.
- exponence, inflectional 14, 18, 20, 26–29, 36f., 41, 45, 48f., 64

- exponential explosion 166, 169f., 178
 extension 6, 8f., 12, 57, 84, 87, 91, 96, 98,
 100, 176f.
- family size effect 3, 128, 133, 215, 217, 233,
 235–238, 244–248
 family types 184
 feedback 5, 84, 88–90, 93, 99, 100
 finite state automaton 140, 149, 151,
 154, 158
 frequency
 token 72, 205–208, 211f., 233
 type 56, 93, 95, 205–207, 212, 252
- gang effects 115f., 165f., 178, 201
 gender 104–106, 110–113, 115f., 123f., 154,
 168, 181
 generative accounts 8–10, 81, 138, 142, 146
 Generalized Context Model 190, 197, 199,
 204f.
 given context 164–167, 169–171, 173, 178
 grammaticality judgments 103, 107, 111
- head families 119–121, 123–136
 hierarchies 171, 173, 174, 216
 ordered 171
 unordered 173
 homogeneity 165f., 170f., 178
 heterogeneity 165f., 169, 176f., 183, 215
- illative
 Estonian 55f.,
 Finnish 180, 182
 Saami 56, 65f.
- infants 102, 108, 111–113, 116
 inference 1, 12, 18, 54, 57, 72f., 80, 129,
 132, 147, 159, 185–187, 192f., 197f.,
 200–202, 208, 210–212
 inflection class 12, 14–20, 32, 36–38, 45,
 48f., 52f., 55, 57, 61, 65, 68–70, 72f., 75,
 79f., 122–124, 188–190, 193, 198, 200,
 204f., 208, 210–212, 215–218, 221–239,
 249f.
- structure of 12, 52, 57, 69, 72, 75, 79f.,
 210–212, 215–217, 234f.
- information theory 56, 150f., 216f., 224, 250
 central concepts 216f.
- innateness 117, 138, 148, 158, 162
 interfix 120–125, 127, 134–136
 implicational structure 11, 56f., 62–64,
 66f., 72, 78
- label morphs 137, 151, 153, 156, 158f.
- learning 11, 18, 25, 41, 50, 53f., 62f., 89,
 101–103, 105f., 110, 112f., 115–117, 122,
 138, 141, 152f., 161f., 184, 194, 251
 leveling 6, 87, 96, 98
 LEXESP corpus 198
 lexical categories 11, 58, 72, 101–104, 106,
 108–110, 113f., 116f.
- lexical decision task 129, 134, 225, 227f.,
 233–235, 237f., 240–248, 250
 lexical processing 215–217, 220f., 223f.,
 226, 229, 233–237, 241, 248, 251f.
- Linguistica* project 147f., 151f.
- measurement, effect of 167
- memory 84–88, 91, 122, 162, 207, 215, 217,
 220, 246, 251
- Minimum Description Length
 (MDL) 140–142, 145f., 148, 150
 modeling 141f.
- Minimal Generalization Learner 95, 193,
 197, 206, 210
- memory 84–88, 91, 122, 162, 166, 207, 215,
 217, 220, 246, 251
- mental lexicon 9f., 61, 81, 88, 135, 215f.,
 224, 250
- modeling 5f., 10f., 95, 115, 122f., 141–143, 164,
 166, 168, 170–172, 174–178, 180, 182, 184f.,
 190, 193, 198, 204, 207, 213, 238, 250
- modifier families 119–124, 126–136
- morphological analyzer 137, 152
 morphological extension, see extension
 morphological family 3, 135, 215f., 232f.,
 235, 237f., 245, 247

- morphological leveling, *see* leveling
- morphophonological patterns 90, 205
- nearest neighbor 165, 176, 178
- Neogrammarian 2
- No-Blur Principle 14, 41, 45, 48f.
- nonce forms 175, 177, 202
- novel words 81, 118, 189f., 198, 200
- outcome 63f., 89, 164f., 170f., 173, 175f., 179–183, 188
 control over selection 85, 89, 176
- overprediction 203
- paradigm
 completion task 11, 102f, 107–111, 113, 115
 inflectional 6, 11, 13–15, 18f., 25–27, 29–41, 45, 49, 50–57, 62f., 65–71, 73–75, 78, 80f., 83, 87, 91, 96–99, 101–103, 105, 107–116, 128, 186f., 209–212, 215, 217f., 221, 223, 227, 232, 234–236, 238–240, 245, 247, 250f.
 information structure of 11f., 52, 56, 62, 64, 67, 69, 72, 75, 79f., 209, 214, 216f., 224, 232
 organization of 57, 74f., 209
 predictability 14, 36–41, 50f.
 uniformity 83, 252
- Paradigm Cell Filling Problem (PCFP) 54–57, 61f., 64–66, 70, 78, 80
- Paradigm Economy Principle 49f.
- paradigmatic relations (*see also* associative relations) 59, 214, 216
- paradigmatic transparency 13–14, 16, 18, 20, 26, 30–36, 38f., 41, 45, 48, 50–53
- past tense
 English 89–91, 118, 176f., 177, 179, 204f., 214, 252
 Finnish 178f.
- pattern
 coherence 100
 extension 57, 84, 87, 100
- phonaesthemes 7
- plurals
 English 4, 87, 177
 Finnish 68f., 72f., 76
 German 123, 183
- principal parts 11, 13, 16–19, 25–36, 38, 41, 45, 51–53, 57, 61f., 69, 209
 dynamic conception 13, 16–19, 25, 31–33, 36, 38, 48, 50f., 69
 static conception 13, 15, -18, 69
- probabilistic approaches 6, 9, 141f., 144, 176, 185f., 188, 190, 192, 194, 196, 198, 200, 204, 206, 208, 210–212, 216
- processing 18, 81, 85–88, 118, 120, 126f., 129f., 134f., 153, 178, 184, 205, 214–217, 220f., 223f., 226, 229, 231–237, 241, 248–252
 latency 215, 225–227, 231, 233–235, 237f., 240–248, 250
- proximity 114–116, 165, 179
- psycholinguistic studies 10, 61, 81, 176, 215f.
- quantum modeling 166, 174
- quantum mechanics 166f.
- redundancy 9, 148f.
- reference 104f., 111, 117, 148
- rule (*vs.* analogy) 9, 10, 81, 90, 118, 120f., 123f., 134, 136, 165–167, 177f., 186, 192–197, 201, 205–207, 210f., 212
 context-sensitive 12, 186, 192–197, 205–207, 210f., 212
- segmentation 8, 138–141, 144, 148, 155
- selection 11, 83–85, 88f., 91–94, 99f., 122f., 127, 129, 134f., 142, 176, 238, 252
 multi-level 84f., 92–94, 100
 by plurality 176
 random 176
- sequential prediction 184

- similarity 1, 2, 6, 84–87, 89–91, 93, 95f.,
 99f., 102, 114f., 122, 153f., 165, 174,
 188–192, 196–198, 200–205, 207f.,
 212f., 217, 242, 248
 structured 190, 192, 196, 201, 213
 variegated 192, 192, 197, 200–204, 207
 simplification 100, 148, 162, 228, 239
 simultaneous prediction 184
 Single Surface Base Hypothesis 50f., 56,
 73, 80
 sound symbolism 7
 Spanish diphthongization 187f., 190,
 192f., 196–202, 204, 206f.
 SPE-style rewrite rules 186, 195, 197
 stress
 Finnish 140, 179f.
 Polish 84
 Spanish 181, 187, 190, 195, 198, 208f., 211
 strings of characters 168
 structural alignment 5f., 114
 structural similarity 1
 suffixes 19, 48f., 58, 83, 105f., 108, 120–123,
 149, 151–154, 156, 205, 236f., 252
 superposition 166f.
 supracontext 114f., 117, 164–166, 168–171,
 173
 syllable structure 173f., 181

 TiMBL (Tilburg Memory-Based
 Learner) 122–124, 136, 193, 205
 traceback method 8
 typological variation 13, 51

 umlaut 123, 183f.
 Universal Grammar 138, 162

 variables
 categorical 165, 169
 dependent 228, 247
 discrete 169, 172
 etymological 179
 independent 64, 67, 164, 166–171, 174

 limits on number of 178f., 184
 scalar variables 169, 172, 181
 semantic 171–173, 181
 unimportant 177f.
 weighting of 180, 182
 variation 3, 10, 13, 51, 55f., 69f., 83–86,
 88f., 91f., 96, 98, 100, 162, 251
 voice onset time 169

 word
 internal structure of 61, 81, 134, 138f.
 segmentation 8, 138–141, 144, 148, 155
 segmentation problem 138–141, 144, 155
 Word and Paradigm (WP) morphology
 56–58.61, 81f., 215f., 235, 251
 wug-test 101, 107, 118

 zeros 168

LANGUAGES

 Archi 55, 94, 122, 162, 171, 173f., 185, 189,
 216

 Catalan 83, 105
 Chinantec, Comaltepec 14, 19f., 25–36,
 38–41, 44f., 51f.
 Chinese 120, 128, 136, 171, 173, 184
 Comaltepec Chinantec (see Chiantec,
 Comaltepec)

 Dutch 120–127, 134–136, 205, 232–235,
 250–252

 English 3f., 6f., 20, 83, 90f., 118–120,
 125–130, 134–136, 144f., 147, 150f., 153,
 155, 168f., 176f., 179, 182, 204f., 214,
 217, 229, 233–237, 243, 250–252
 English, American 85, 87
 English, British 237
 English, Old 89
 Estonian 55, 57, 61

- Finnish 56, 61, 68–73, 76, 140, 178–180, 182
French 7, 119f., 130, 132, 154–156, 205
Fur 14, 48–52
- German 120, 123f., 134–136, 140, 179, 183,
205, 209
Georgian 55, 59
Greek 209
- Indonesian 120, 130, 135
- Japanese 120, 124, 134–136
- Korean 209
- Latin 13–15, 18, 143, 179, 182, 234
- Nenets, Tundra 56, 58f., 61, 72–80
- Polish 48f., 84
Portuguese 146f.
- Russian 106, 108–116
- Saami (see Saami, Northern)
Saami, Northern 56, 61, 65–69, 73, 76
Serbian 217–219, 221–224, 234-f.,
249–251
Spanish 154, 181, 187–190, 192, 195f., 198,
200f., 204, 206, 208–212
Swahili 138, 154, 157–159, 161
- Tigrinya 6
- Uralic 61, 64, 72
- Yupik 3